HIGH-THROUGHPUT CHARACTERIZATION OF FOODBORNE PATHOGENS USING NEXT-GENERATION SEQUENCING

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Laura M. Carroll

August 2019

HIGH-THROUGHPUT CHARACTERIZATION OF FOODBORNE PATHOGENS USING NEXT-GENERATION SEQUENCING

Laura M. Carroll, Ph.D.

Cornell University 2019

Next-generation sequencing (NGS) is being increasingly employed to characterize food-associated microbes and communities, including those which pose a threat to human health. As the amount of publicly available genomic data from these organisms increases, (i) rapid, scalable methods for inferring biological function from large amounts of NGS data are needed, and (ii) meaningful biological conclusions derived using these methods can be leveraged to improve safety along the food supply chain. The studies reported here detail the application of whole-genome sequencing (WGS) to two groups of organisms which differ in terms of the challenges they pose to human health: (i) non-typhoidal *Salmonella enterica*, a well-characterized, Gram-negative foodborne pathogen which boasts a large repertoire of established computational methods for analyzing WGS data derived from it, and (ii) the lesser-sequenced *Bacillus cereus* group, which consists of closely related, Gram-positive, spore-forming species which vary in their ability to cause disease in humans.

For *Salmonella enterica*, antimicrobial resistance (AMR) was of particular concern; WGS was used to characterize 90 AMR strains isolated from either human or bovine hosts from New York or Washington State. In addition to predicting phenotypic resistance to a panel of twelve antimicrobials with high accuracy (mean sensitivity and specificity of 97.2% and 85.2%., respectively), *in silico* characterization of AMR determinants present in all isolates unveiled

significant geographic and host associations, including quinolone resistance, which was only observed in human isolates from Washington State. Additionally, one multidrug-resistant, colistin-susceptible *Salmonella* Typhimurium strain was found to harbor *mcr-9*, a novel plasmid-mediated colistin resistance gene.

For *Bacillus cereus*, classification of isolates based on virulence potential was the primary focus. An *in silico* typing tool designed to rapidly identify *B. cereus* group virulence factors and taxonomic affiliation using WGS data is described. This application, named BTyper, was used to query all *Bacillus cereus* group genomes submitted to NCBI's Genbank database ($n = 662$, accessed April 6, 2017). Additionally, BTyper was used to characterize the genomes of 33 *B. cereus* group strains isolated in conjunction with a 2016 outbreak. Thirty genomes were classified as emetic *Bacillus cereus* and predicted to be the cause of a single-source outbreak using a combination of computational, microbiological, and epidemiological methods.

Overall, the results presented here showcase how NGS can be used to characterize food-associated microbes at greater resolution than preceding technologies. Additionally, computational and statistical methods used to analyze Illumina data derived from foodborne pathogens are emphasized. The tools and methods detailed here can serve as a guide for deriving biologically informed conclusions from WGS data.

BIOGRAPHICAL SKETCH

Laura M. Carroll grew up in Houghton, Michigan. She attended Michigan State University from 2009 to 2014, where she received a Professorial Assistantship through the university's Honors College to conduct research under the direction of Professor Brad Marks. As a member of the Biosystems Engineering Food Safety Laboratory, Laura spent five years developing mathematical models to predict the thermal inactivation of foodborne pathogens in various food matrices, with an emphasis on modeling the physiological response of *Salmonella enterica* to prolonged periods of sublethal thermal stress.

After graduating with a B.S. in Genomics and Molecular Genetics and a B.A. in History, Laura began her graduate studies at Cornell University under the direction of Professor Martin Wiedmann. As a doctoral student, Laura's research focused on (i) developing bioinformatic pipelines to rapidly characterize bacteria *in silico* using next-generation sequencing data, and (ii) using those pipelines to analyze large genomic data sets from bacterial isolates and microbial communities. During her time at Cornell, Laura received a National Science Foundation (NSF) Graduate Research Fellowship, and, later, a NSF Graduate Research Opportunities Worldwide (NSF GROW) award, which allowed her to spend time as a visiting researcher with Professor Tanja Stadler's Computational Evolution Group at ETH Zurich in Switzerland. She additionally spent several months as a graduate intern with IBM's Industrial and Applied Genomics Group, where she was first introduced to metagenomic and metatranscriptomic data analysis methods. After completing her Ph.D., Laura will be focusing primarily on metagenomic and metatranscriptomic data analysis as a Postdoctoral Fellow in the group of Dr. Georg Zeller at the European Molecular Biology Laboratory (EMBL) in Heidelberg, Germany.

To my parents, for their unwavering love and support

ACKNOWLEDGEMENTS

It is impossible to allocate the space necessary to adequately thank all of those who have helped me reach this point in my career. I am indebted to my committee members, Drs. James Booth and Michael Stanhope, for their guidance and mentorship, as well as the National Science Foundation Graduate Research Fellowship Program (NSF GRFP) and associated NSF Graduate Research Opportunities Worldwide (NSF GROW) program for their generous funding.

I would not be here, in the most literal sense, without the support of my family: my mother, who, as a woman in STEM, has been a role model available to me for the entirety of my life; my father, for his unconditional love and support, even when I pushed the boundaries of "unconditional"; my brother, for his willingness to drop everything to help me, even when I probably (read: definitely) don't deserve it; and my sister, who has been, and will forever be, my confidant, favorite labmate, and best friend.

Professionally, I am beholden to my undergraduate research advisor, Dr. Brad Marks, for the essential mentorship he provided while I navigated my undergraduate years and transition to graduate school; Nicole Hall, who eventually molded me into a semi-functioning member of a laboratory; Dr. Teresa Bergholz, whose guidance (and patience) nurtured my love of research (but not RNA); Dr. Henk den Bakker, who helped me hit the ground running in my first few weeks of graduate school (and continues to help me, even when I pester him from afar); Dr. Richard Pereira, who guided me through my first research project at Cornell; Drs. Simone Bianco and Kristen Beck, whose mentorship during my time at IBM fostered my love of shotgun metagenomic and metatranscriptomic data analysis; Dr. Ahmed Gaballa, who is possibly the only person on the planet as enthusiastic about colistin resistance as I am; and Dr. Tanja

Stadler, who welcomed me into her group and made my time in Switzerland easily one of the most transformative experiences I have had as a graduate student and researcher.

I am especially grateful for the guidance I have received from Drs. Jasna Kovac (my "*Bacillus* advisor"), Claudia Guldimann (my "*Salmonella* advisor"), and Rachel Cheng, who have been incredible mentors, role models, collaborators, and friends throughout my graduate career. I consider myself incredibly fortunate to be able to work alongside such brilliant researchers who display such an aspirational work ethic and level of scientific creativity.

Continuing on a personal level, I am indebted to all of those on whom I have leaned at various times during my graduate career and beyond: my soulmates, Rachel Allison, Geoff Pleiss, and Tobias Schnabel; my sisters, Corinna Noel and Jillian Jastrzembski; my "sisters", Ariel Buehler and Lory Henderson; moja sestra, Svetlana Lyalina; my Swiss-ters, Jana Huisman, Rachel Warnock, Joelle Barido-Sottani, and Julia Pecherska; Venelin Mitov and Daniel Scain Farenzena, who went out of their way to make Basel feel like my second home; and to all those who have been there for me in more ways than they can possibly know: Pedro Menchik; Bryan Peele; Madeleine Bee; Emily Griep; Jeff Tokman; Sophia Harrand; Beth Burzynski; Gorjan Dukovski; Richard Goater; Veronica Guariglia; Dave Kent; Vlad Niculae; Madelyn Shoup; Hilary Podgers; Brittany Massa; Morgan Frost; Kylie Gignac; Ian Hildebrandt; Dani Smith; and Sarah Buchholz.

I owe additional gratitude to all of my labmates, past and present, whom I was unable to list here, particularly my friends and colleagues in the Biosystems Engineering Food Safety Laboratory at Michigan State University, IBM's Industrial and Applied Genomics Group, the Computational Evolution Group

at ETH Zurich, and the Food Safety Laboratory and Milk Quality Improvement Program at Cornell University.

Finally, I would like to thank my advisor of the past five years, Dr. Martin Wiedmann. Articulating how grateful I am to have him as a mentor is completely futile; the level of independence and flexibility he has afforded me as a doctoral student to pursue nearly every research question that I could dream up is incomparable (and, as he would probably argue, excessive). I consider myself infinitely fortunate to have been a member of his laboratory, and I will never take the knowledge, skills, and lessons he has taught me, both as a researcher and as a person, for granted.

TABLE OF CONTENTS

[1] From Wiedmann, Martin and Laura M. Carroll (2019). "Next-Generation Sequencing". In: *Encyclopedia of Food Chemistry* , pp. 376-383. DOI: 10.1016/b978-0-08-100596-5.21792-7.

[2] From Carroll, Laura M., Martin Wiedmann, Henk den Bakker, Julie Siler, Steven Warchocki, David Kent, Svetlana Lyalina, Margaret Davis, William Sischo, Thomas Besser, Lorin D. Warnick, and Richard V. Pereira (2017). "Whole-Genome Sequencing of Drug-Resistant *Salmonella enterica* Isolates from Dairy Cattle and Humans in New York and Washington States Reveals Source and Geographic Associations". In: *Applied and Environmental Microbiology* 83, pp. e00140-17. DOI: https://doi.org/10.1128/AEM.00140-17.

[3]From Carroll, Laura M., Ahmed Gaballa, Claudia Guldimann, Genevieve Sullivan, Lory O. Henderson, and Martin Wiedmann (2019). "Identification of Novel Mobilized Colistin Resistance Gene *mcr-9* in a Multidrug-Resistant, Colistin-Susceptible *Salmonella enterica* Serotype Typhimurium Isolate". In: *mBio* 10, pp. e00853-19. DOI: 10.1128/mBio.00853-19.

[4]From Carroll, Laura M., Jasna Kovac, Rachel A. Miller, and Martin Wiedmann (2017). "Rapid, High-Throughput Identification of Anthrax-Causing and Emetic *Bacillus cereus* Group Genome Assemblies via BTyper, a Computational Tool for Virulence-Based Classification of *Bacillus cereus* Group Isolates by Using Nucleotide Sequencing Data". In: *Applied and Environmental Microbiology* 83, pp. e01096-17. DOI: 10.1128/AEM.01096-17.

[5] From Carroll, Laura M., Martin Wiedmann, Manjari Mukherjee, David C. Nicholas, Lisa A. Mingle, Nellie B. Dumas, Jocelyn A. Cole, and Jasna Kovac (2019). "Characterization of Emetic and Diarrheal *Bacillus cereus* Strains From a 2016 Foodborne Outbreak Using Whole-Genome Sequencing: Addressing the Microbiological, Epidemiological, and Bioinformatic Challenges". In: *Frontiers in Microbiology* 10, pp. 144. DOI: 10.3389/fmicb.2019.00144.

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION[1]

[1]FROM WIEDMANN, MARTIN AND LAURA M. CARROLL (2019). "NEXT-GENERATION SEQUENCING". IN: *ENCYCLOPEDIA OF FOOD CHEMISTRY* , PP. 376-383. DOI: 10.1016/B978-0-08-100596-5.21792-7.

1.1 Next-Generation Sequencing: an Overview

Next-generation sequencing (NGS) encompasses sequencing technologies that are capable of sequencing many DNA strands in parallel, resulting in higher throughput than can be achieved using Sanger sequencing. As NGS has become cheaper and more accessible, it has been used to address an expanding range of biological problems, including many relevant to food safety and quality.

Contemporary NGS sequencing platforms employ either a (i) short-read, or (ii) long-read sequencing approach (Table 1.1). Short-read sequencing approaches typically yield read lengths of up to 700 base pairs (bp), which tend to be shorter than those produced by Sanger sequencing (Goodwin, McPherson, and McCombie 2016; Liu et al. 2012). Currently, sequencing-by-synthesis approaches (SBS) to NGS are the dominant paradigm in short-read sequencing. These approaches (e.g. Illumina sequencing, Roche 454 pyrosequencing, Ion Torrent semiconductor-based sequencing) rely on the use of DNA polymerase in their respective methods (Goodwin, McPherson, and McCombie 2016). SBS approaches to short-read sequencing can be contrasted with the sequencing-by-ligation (SBL) approach employed by the SOLiD (Small Oligonucleotide Ligation and Detection) platform, which employs DNA ligase to join fluorescently-labelled probe and anchor sequences to a DNA strand (Goodwin, McPherson, and McCombie 2016). Among the SBS approaches and short-read sequencing methods as a whole, Illumina sequencing has emerged as the dominant technology (Goodwin, McPherson, and McCombie 2016), in which fluorescently-tagged nucleotides are added in complement to amplified strands of DNA. Upon the addition of a single nucleotide, the fluorescent dye is imaged, and the identity of the corresponding base is recorded (Goodwin, McPherson, and

McCombie 2016).

Table 1.1: Overview of next-generation sequencing technologies discussed in this chapter.[a]

Sequencing technology	*Sequencing mechanism*	*Read length*[b]	*Error rate (type of error)*
Sequencing-by-ligation (SBL)			
SOLiD	Ligation; 2-base encoding	50-75 bp	≤ 0.1% (AT bias)[c]
Sequencing-by-synthesis (SBS)			
454	Pyrosequencing	Up to 1000 bp	1% (indel)[d]
Illumina	Illumina SBS	25-300 bp; can be 100 Kb if synthetic long-read library preparation is used	0.1% to 1%, depending on platform/output (substitution)
Ion Torrent	Hydrogen ion detection	Up to 400 bp	1% (indel)
Single-molecule long-read			
Oxford Nanopore	Nanopore	Up to 200 Kb	12% (indel)
Pacific Biosciences	Single-molecule real-time sequencing	8-20 Kb	13% for a single pass (indel)

[a]Summarized from reviews of NGS technologies by Goodwin et al., Liu, et al., and Glenn (Goodwin, McPherson, and McCombie 2016; Liu et al. 2012; Glenn 2011)
[b]bp, base pairs; Kb, kilobase pairs.
[c]AT, adenine and thymine.
[d]indel, insertion/deletion.

While short-read sequencing technologies have been the workhorse of NGS, they are not without limitations; many genomic features, such as long, repetitive regions or copy number variations, cannot be readily resolved using short reads (Goodwin, McPherson, and McCombie 2016). Long-read sequencing technologies have been able to bridge the literal gaps that their short-read counterparts have been unable to resolve, relying on either (i) synthetic long-read approaches or (ii) single-molecule long-read sequencing approaches (Pacific Biosciences and Oxford Nanopore) (Goodwin, McPherson, and McCombie 2016). Synthetic long-read sequencing approaches employ existing short-read sequencing platforms, but use barcoding during library preparation to link fragments (Goodwin, McPherson, and McCombie 2016). Single-molecule long-read sequencing approaches, however, yield "true" long reads that can span kilobases, with the approach most commonly employed as of late 2017 being the Pacific Biosciences (PacBio) single-molecule real-time (SMRT) approach (Goodwin, McPherson, and McCombie 2016). SMRT sequencing uses a DNA polymerase fixed to the bottom of a well in a specialized flow cell through which a DNA strand is passed (Goodwin, McPherson, and McCombie 2016). Upon the incorpora-

tion of a single, fluorescently-labelled nucleotide by the polymerase, light is emitted and recorded by a camera to determine the identity of the nucleotide (Goodwin, McPherson, and McCombie 2016). This can be contrasted with the aforementioned short-read SBS approaches, which rely on DNA polymerase traversing the DNA template to which it is bound (Goodwin, McPherson, and McCombie 2016). In addition to the PacBio platform, the small and highly portable MinION platform from Oxford Nanopore Technologies also employs a single-molecule long-read sequencing approach, during which a strand of DNA is passed through a protein pore along with an electric current (Goodwin, McPherson, and McCombie 2016). As different combinations of nucleotides are passed through the pore, shifts in the electric current are recorded (Goodwin, McPherson, and McCombie 2016).

Long-read sequencing is becoming increasingly popular for many applications, including gap closure in reference genomes, characterization of long genomic structures, and the generation of closed chromosomes or transcriptomes (Goodwin, McPherson, and McCombie 2016). A notable consideration when comparing short-read and long-read sequencing methods is the relatively high error rates of long-read sequencing platforms (Goodwin, McPherson, and McCombie 2016). For example, the PacBio RS II, which yields average read lengths of 10-15 Kb, has an error rate as high as 15% for a single pass through a molecule of DNA (Goodwin, McPherson, and McCombie 2016). However, this error rate can be reduced to one that rivals that of Sanger sequencing by increasing sequencing coverage through multiple passes; after 30 passes (i.e. at 30X coverage), the accuracy of the consensus is greater than 99.999% (http://www.pacb.com/smrt-science/smrt-sequencing/accuracy/ and https://www.pacb.com/uncategorized/a-closer-

look-at-accuracy-in-pacbio/) (Goodwin, McPherson, and McCombie 2016).

1.2 NGS Data Analysis

Processing and analysis of NGS data is dependent on the sequencing technology used, as well as the experimental goals. Regardless of sequencing method or experimental design, the first steps in the analysis of NGS data usually involve an assessment of read quality, using metrics such as the total number of reads, the distribution of read lengths, sequence quality scores, etc. This can be followed by trimming of adapters and/or low-quality bases, filtering out low-quality reads, and filtering of contaminant DNA, steps for which a number of programs are available (Breitwieser, Lu, and Salzberg 2017). After these pre processing steps, data analysis can be carried out according to the goals of the experiment, with possible food science-relevant applications discussed below (Table 1.2).

Table 1.2: Overview of food science-relevant next-generation sequencing applications discussed in this chapter.

Next-generation sequencing application	*Number of organisms queried*	*Nucleic acid extracted/sequenced*	*Genomic elements queried*	*Current food science-relevant applications*
Whole-genome sequencing (WGS)	1	DNA/DNA	Entire genome	Characterization of food-relevant organisms at the genomic level
RNA sequencing (RNA-Seq)	1	RNA/cDNA reverse-transcribed from RNA	Entire transcriptome	Characterization of food-relevant organisms at the transcriptional level
High-throughput amplicon sequencing (e.g. 16S rDNA sequencing, DNA-barcoding)	≥ 1	DNA/DNA	Selected amplicon(s) present in sample (usually 16S rDNA for bacterial/archaeal communities; other loci for eukarya)	Taxonomic characterization of food-relevant microbial communities (usually bacterial/archaeal communities); authentication of eukaryotic food matrices (e.g. seafood, meat products)
Shotgun metagenomic sequencing	> 1	DNA/DNA	All genomes present in sample	Characterization of food-relevant communities at the genomic level (queries eukarya, bacteria, archaea, and viruses)
Shotgun metatranscriptomic sequencing	> 1	RNA/cDNA reverse-transcribed from RNA	All transcriptomes present in sample	Characterization of food-relevant communities at the transcriptional level (queries eukarya, bacteria, archaea, and viruses)

1.3 NGS Applications: Whole-Genome Sequencing of Microbial Contaminants

Traditionally, microbial contaminants isolated from food undergo various organism-specific phenotypic or biochemical tests (e.g. testing for motility, toxin production, growth at various temperatures) to elucidate or confirm their identity (FDA 1998). These tests may be supplemented with additional typing methods, such as serotyping, pulsed-field gel electrophoresis (PFGE), Sanger sequencing of a single taxonomic marker gene or genomic region (i.e. single-locus sequence typing; SLST), or Sanger sequencing of multiple loci used in a multi-locus sequence typing (MLST) scheme (Kovac et al. 2017; Sabat et al. 2013). However, the per-isolate cost of whole-genome sequencing (WGS) has decreased to the point at which it is comparable, and even below, the price of many of these traditional subtyping methods (Kovac et al. 2017), making it an increasingly popular method for characterizing microbial contaminants isolated from food matrices, food-associated environments (e.g farm environments, processing environments), and, in the case of pathogenic microbes, from hosts (e.g. in human- or animal-clinical settings) (Kovac et al. 2017). Furthermore, many of these typing methods (e.g. serotyping, SLST, MLST) can be performed *in silico* using WGS data, with the advantage that one can query the majority of a microbial genome from a single data set, rather than just a small fraction of it ($< 0.01\%$ for a traditional 7-gene MLST scheme) (Kovac et al. 2017). In addition to *in silico* subtyping, WGS data from microbial contaminants can be used to predict functional characteristics of isolates, query genes or genomic elements of interest within a genome (e.g. plasmids, bacteriophage, and genes contributing to antimicrobial resistance or virulence), and, in the case of pathogenic microor-

ganisms, detect and track outbreaks (Kovac et al. 2017).

After sequencing the genomic DNA and pre-processing the resulting reads from a microbial isolate (see "NGS Data Analysis" section above), possible analysis steps that may be taken include (i) *de novo* genome assembly of the reads into contiguous stretches of sequence (contigs) (Giordano et al. 2017; Liao, S.-H. Lin, and H.-H. Lin 2015; Ekblom and Wolf 2014), (ii) mapping reads back to a reference genome, (iii) identifying single-nucleotide polymorphisms (SNPs), insertions, and deletions (indels) in NGS data through variant calling (Olson et al. 2015), (iv) constructing phylogenetic trees to assess the evolutionary relationship of multiple isolates, (v) assigning allelic types at a genomic scale using core genome or whole genome multi-locus sequence typing (cgMLST and wgMLST, respectively), and (vi) locating genes and features in NGS data via genome annotation (Richardson and Watson 2012; Mudge and Harrow 2016; Yandell and Ence 2012). These data can be used to characterize isolates at high resolution, making it possible to compare isolates geospatially and temporally at the whole-genome scale.

WGS is becoming an increasingly valuable tool for characterizing microbial contaminants, particularly pathogens, isolated from food and food processing environments. A notable example of the utility of WGS can be seen in the multi-agency collaboration in the US to sequence all *Listeria monocytogenes* isolates from human patients, food, and the environment (Jackson et al. 2016). Since its implementation in 2013, the WGS-based surveillance program detected more listeriosis clusters and solved more outbreaks each year, relative to the previous year (Jackson et al. 2016). Similar findings have been seen for *Salmonella enterica* serotype Enteritidis (*S.* Enteritidis); retrospective sequencing of 55 *S.* En-

teritidis from clinical and environmental sources allowed isolates from known outbreaks to be differentiated from sporadic isolates at greater resolution than PFGE (Taylor et al. 2015). These examples showcase how WGS can be used to not only characterize foodborne pathogens at high resolution, but also the outbreaks associated with them.

1.4 NGS Applications: RNA Sequencing (RNA-Seq) of Food-Relevant Organisms

While WGS can be used to characterize the genome of an organism at unprecedented resolution, it offers no information on whether a genomic element of interest is being actively transcribed or not. This is particularly important within a food safety context; for example, the mere isolation of a pathogen from a food matrix does not necessarily mean that particular isolate is viable, or that it is transcribing the genes necessary to cause infection or intoxication in a human host. Traditionally, quantitative reverse-transcription PCR (RT-qPCR) has been employed to quantify or detect shifts in transcript levels of loci of interest. For this method, reverse-transcription PCR (RT-PCR) is used to obtain complementary DNA (cDNA) from a RNA template, and the resulting cDNA can be quantified using quantitative PCR (qPCR). In a food science context, RT-qPCR has been proposed as a method for detecting viable microorganisms, quantifying virulence, toxin, or stress response gene transcription, and quantifying microbial growth in food matrices (Postollec et al. 2011; Carroll et al. 2016). Studying transcription at a genome-wide scale, however, was made possible with cDNA microarrays, which have been used to study the stress responses of various

foodborne pathogens, as well as their transcription of toxin and virulence genes (Postollec et al. 2011; Roy and Sen 2006; Rasooly and Herold 2008). As NGS has become more feasible, however, it is now possible to query the transcriptome of an organism in its entirety at low cost: RNA sequencing (RNA-Seq) employs NGS technologies to sequence cDNA reverse-transcribed from RNA that has been extracted from an organism of interest (Z. Wang, Gerstein, and Snyder 2009). RNA-Seq allows one to quantitatively survey transcribed regions of an entire genome, improving upon microarrays in both cost and flexibility (i.e. the ability to characterize any organism that can be sequenced, rather than relying on the availability of an array for a particular organism), which is particularly valuable for studying organisms or genomic regions that may not be well-characterized.

After employing NGS to sequence cDNA from an organism of interest, and determining that the quality of sequencing reads is adequate, reads are usually aligned to a reference genome or an assembled transcriptome (McClure et al. 2013; Conesa et al. 2016). After assessing mapping quality and determining that it is appropriate, reads mapping to various genes or genomic regions can be quantified and normalized, taking into account biases such as gene length (McClure et al. 2013; Conesa et al. 2016). After quantification and normalization, analyses can be carried out according to the experimental goals (e.g. differential transcription under various conditions). Within the realm of food safety, RNA-Seq has been applied to pathogenic and toxin-producing microorganisms to identify differentially-transcribed genes during growth in various food matrices (Tang et al. 2015; Deng, Z. Li, and W. Zhang 2012; Galia et al. 2017), after exposure to various stressors (e.g. acid, starvation, or antimicrobial stressors) (F. Zhang et al. 2014; Casey et al. 2014; Butcher and Stintzi 2013; K. Jia et al.

2017), and during the infection of a host (Avraham et al. 2016).

1.5 NGS Applications: High-Throughput Amplicon Sequencing

WGS and RNA-Seq have allowed food-associated microorganisms to be characterized at unprecedented resolution. However, these methods typically require the microorganism in question to be in pure culture or isolated via culture-based methods, a process which involves the use of organism-specific enrichment media, selective media, and isolation protocols (Kovac et al. 2017). Metagenomics, which involves sequencing DNA directly from an environmental sample, attempts to bypass the isolation step, making it possible to survey an entire community simultaneously (Kovac et al. 2017).

Until recently, NGS-based metagenomic methods have primarily involved high-throughput amplicon sequencing. Also referred to as ”metataxonomics”, ”meta-genetics”, or ”marker gene metagenomics”, high-throughput amplicon sequencing employs NGS technologies to sequence targeted PCR products (amplicons) to characterize particular communities. When surveying bacterial and archaeal communities, the 16S ribosomal DNA gene (16S rDNA) is usually the amplicon of choice, as it is present in all bacterial and archaeal species. 16S rDNA sequencing has been used to survey the microbiota of various foods (De Filippis, Parente, and Ercolini 2016; Kergourlay et al. 2015; Ercolini 2013), including fermented foods (De Filippis, Parente, and Ercolini 2016) and food matrices subjected to pathogen-specific enrichments (Jarvis et al. 2015; Lusk et al. 2012), as well as to monitor bacterial community shifts in food processing

environments (Stellato et al. 2016; Hultman et al. 2015).

One of the strengths of 16S rDNA amplicon sequencing is that there are many freely available bioinformatic tools and pipelines available for data analysis and visualization of results (e.g. QIIME, Mothur). A typical workflow for analyzing NGS data from high-throughput 16S rDNA experiments may include pre-processing of the raw reads, clustering of sequences into operational taxonomic units (OTUs) based on sequence similarity, and taxonomic assignment of sequences using a database of 16S rDNA genes (e.g. RDP, Greengenes, SILVA) (Oulas et al. 2015; Siegwald et al. 2017).

In addition to querying bacterial and archaeal communities, the same principals of amplicon sequencing can be applied to characterize eukarya. DNA-barcoding, a practice in which a specific region of a genome is sequenced, is a commonly-used method for food matrix authentication along the food supply chain (Ellis et al. 2016; Galimberti et al. 2013). For this approach, a genetic marker (i.e. a "barcode") present in a range of taxa, but variable enough to be capable of discriminating between taxa of interest, is sequenced (Galimberti et al. 2013), similar to the way the 16S rDNA gene is used to survey bacterial/archaeal communities. When querying animal DNA in a matrix (e.g. for seafood or meat authentication), the cytochrome b (*cytB*) and cytochrome c oxidase subunit 1 (*COI*) genes are common amplicons of choice. For fungi, the internal transcribed spacer (ITS) region of the genome is the locus of choice (Schoch et al. 2012), while a number of loci have been proposed for querying plant DNA present in a matrix (Hollingsworth, Graham, and Little 2011; Hollingsworth, D.-Z. Li, et al. 2016). The sequences of these genes are then compared to the barcodes of known taxa, such as those found in the Barcode of Life Database

(BOLD) (Ratnasingham and Hebert 2007) or the National Center for Biotechnology Information's (NCBI) GenBank database (Benson et al. 2013). Applications of DNA-barcoding within the realm of matrix authentication and contaminant detection along the food supply chain have included authentication of and contaminant detection in seafood (Carvalho, Palhares, Drummond, and Frigo 2015; Armani et al. 2015; Pardo, Jimenez, and Perez-Villarreal 2016; Kim et al. 2015; Chang et al. 2016; Carvalho, Palhares, Drummond, and Gadanho 2017), meat (Kane and Hellberg 2016; Hellberg, B. C. Hernandez, and E. L. Hernandez 2017; Naaum et al. 2018), poultry (Hellberg, B. C. Hernandez, and E. L. Hernandez 2017), dairy products (Galimberti et al. 2013), olive oil (Kumar, Kahlon, and Chaudhary 2011), and spices (Swetha et al. 2017; De Mattia et al. 2011; Galimberti et al. 2013).

Until recently, DNA-barcoding was limited by the low-throughput that Sanger sequencing provides; however, NGS has emerged as a low-cost, high-throughput alternative (Ellis et al. 2016; Shokralla et al. 2014) that has been used for characterizing both raw ingredients and processed foods (Galimberti et al. 2013). In this high-throughput approach, sequencing reads are mapped to sequences in an appropriate database (often BOLD or GenBank) after determining that read quality is appropriate. The proportion of reads mapping to a particular species in the database corresponds to the proportion of that particular species in the matrix. A notable example of the application of high-throughput sequencing for food matrix authentication is provided by Carvalho et al. (Carvalho, Palhares, Drummond, and Gadanho 2017), in which mislabeled cod products in Brazilian stores and restaurants were identified by targeted sequencing of the *cytB* and *COI* genes present in processed cod products using NGS (Carvalho, Palhares, Drummond, and Gadanho 2017). In addition

to identifying mislabeled products, the composition of blended products composed of multiple fish species could be determined by sequencing the selected loci (Carvalho, Palhares, Drummond, and Gadanho 2017).

1.6 NGS Applications: Shotgun Metagenomic and Metatranscriptomic Sequencing

Although high-throughput amplicon sequencing has offered a higher-resolution glimpse into food and food-associated microbiomes, it has numerous limitations that are particularly relevant within the realms of food safety and food quality, perhaps most notably the inability to query organisms that do not possess the amplicon of choice (e.g. eukarya in a community cannot be queried if 16S rDNA amplicon sequencing is performed; see "NGS Applications: High-Throughput Amplicon Sequencing" section above). For 16S rDNA amplicon sequencing of bacterial/archaeal communities, additional drawbacks include (i) difficulty achieving species-level resolution (Janda and Abbott 2007; Rossi-Tamisier et al. 2015) and reliably distinguishing pathogenic bacteria from non-pathogenic species (e.g. *L. monocytogenes* from *Listeria innocua*, human pathogens *Bacillus anthracis* from *Bacillus cereus* and biopesticide *Bacillus thuringiensis*), (ii) PCR amplification and primer bias (Brooks et al. 2015), and (iii) inability to query functionally-relevant genomic elements directly, such as virulence or antimicrobial resistance determinants (Kovac et al. 2017).

An increasingly popular alternative to amplicon sequencing is shotgun metagenomic sequencing, an approach in which all DNA present in a sample is sequenced, rather than solely an amplicon. By sequencing all DNA present in

a sample, the amplification bias and low taxonomic and functional resolution issues which plague amplicon sequencing can typically be bypassed (Kovac et al. 2017). In addition to sequencing all of the bacterial and archaeal DNA present in a sample, all viral and eukaryotic DNA is sequenced; this is particularly relevant when the community of interest is derived from a eukaryotic matrix (e.g. from a host or from food), as the majority (as much as 99%) of DNA will come from the eukaryotic matrix itself (Kovac et al. 2017; Noyes et al. 2016). While large quantities of host DNA may not be a problem if the experimental goal is to assess the composition of the food matrix itself, it may hinder the sequencing and detection of many microbial species. As a result, when extracting DNA from a matrix containing high amounts of host DNA, additional steps may be taken to deplete any background DNA originating from the matrix itself to increase the proportion of microbial DNA that is sequenced (Kovac et al. 2017). After sequencing the extracted DNA, analysis of the resulting sequencing reads is carried out according to the experimental goals, which may include taxonomic assignment (Sharpton 2014), metagenomic assembly, functional annotation (Sharpton 2014), and/or conducting a metagenome-wide association study by associating community data with a particular phenotype (J. Wang and H. Jia 2016; Lynch and Pedersen 2016).

As with all genomic approaches, shotgun metagenomic methods can offer insight into the genomic composition of a community, but cannot offer information as to which genes are being transcribed and possibly translated and expressed as protein products (Kovac et al. 2017). Similar to the way RNA-Seq can be used to complement WGS of a bacterial isolate, metagenomic approaches can be supplemented with shotgun metatranscriptomic sequencing, which involves sequencing cDNA reverse-transcribed from RNA (typically messenger

RNA) extracted from an entire community (Kovac et al. 2017).

Analysis of shotgun metagenomic and metatranscriptomic data usually begins with pre-processing steps such as assessing read quality and trimming adapters (Breitwieser, Lu, and Salzberg 2017). This can be followed by (i) assembly of the reads into contigs, or (ii) taxonomic or functional classification directly from sequencing reads (Breitwieser, Lu, and Salzberg 2017). For a review of methods for metagenomic data analysis, see Breitwieser et al. (Breitwieser, Lu, and Salzberg 2017).

The use of shotgun metagenomic and metatranscriptomic approaches to survey communities in foods has been undertaken only recently (De Filippis, Parente, and Ercolini 2016). Goals of these studies have included characterization of the microbiomes of various foods in the presence of foodborne pathogens and/or spoilage organisms (Jarvis et al. 2015; Ottesen et al. 2013), tracking foodborne pathogens and antimicrobial resistance genes along the food supply chain (Noyes et al. 2016; Yang et al. 2016), characterizing eukaryotic food matrices composed of multiple species (Ripp et al. 2014), and characterizing the microbiomes of various food matrices during processes such as fermentation (De Filippis, Parente, and Ercolini 2016; Kergourlay et al. 2015; Alkema et al. 2016; Valdes et al. 2013; Lessard et al. 2014; De Filippis, Genovese, et al. 2016; Monnet et al. 2016). A notable example of the application of shotgun metaomics approaches to identify the cause of a food quality anomaly is provided by Quigley et al. (Quigley et al. 2016); using high-throughput 16S rDNA sequencing followed by shotgun metagenomic sequencing, *Thermus thermophilus* was proposed (and later confirmed) to be the cause of a pink discoloration defect in Continental-type cheeses (Quigley et al. 2016).

1.7 Conclusion

NGS technologies are being employed increasingly in food science relevant realms, with applications ranging from surveying microbial communities involved in food processing, to rapidly characterizing bacterial isolates from foodborne outbreaks. As sequencing costs continue to decrease, it is likely that whole-genome and meta-omics approaches will be applied routinely at various points along the food supply chain.

The following chapters detail how NGS can be used to query bacterial foodborne pathogens, with an emphasis on rapid, high-throughput computational methods which can be used to analyze short-read data produced by Illumina platforms. Two model organisms are discussed: (i) non-typhoidal *Salmonella enterica*, a widely studied Gram-negative pathogen which can be transmitted between livestock and humans, as well as though food, and (ii) the lesser-queried Gram-positive members of the *Bacillus cereus* group, which are spore-forming organisms commonly isolated from soil. While both groups of organisms are capable of causing foodborne illness in humans, they differ at a biological level and, thus, necessitate different approaches to analyze NGS data derived from them.

1.8 References

Alkema, Wynand, Jos Boekhorst, Michiel Wels, and Sacha A. F. T. van Hijum (2016). "Microbial bioinformatics for food safety and production". In: *Brief Bioinform* 17.2, pp. 283–292. DOI: 10.1093/bib/bbv034.

Armani, A. et al. (2015). "DNA barcoding reveals commercial and health issues in ethnic seafood sold on the Italian market". In: *Food Control* 55, pp. 206–214.

Avraham, Roi et al. (2016). "A highly multiplexed and sensitive RNA-seq protocol for simultaneous analysis of host and pathogen transcriptomes". In: *Nature Protocols* 11, pp. 1477–1491.

Benson, Dennis A. et al. (2013). "GenBank". In: *Nucleic Acids Res* 41.Database issue, pp. D36–D42. DOI: 10.1093/nar/gks1195.

Breitwieser, Florian P., Jennifer Lu, and Steven L. Salzberg (2017). "A review of methods and databases for metagenomic classification and assembly". In: *Briefings in Bioinformatics*. DOI: 10.1093/bib/bbx120. eprint: http://oup.prod.sis.lan/bib/advance-article-pdf/doi/10.1093/bib/bbx120/20139928/bbx120.pdf.

Brooks, J. Paul et al. (2015). "The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies". In: *BMC Microbiol* 15, pp. 66–66. DOI: 10.1186/s12866-015-0351-6.

Butcher, James and Alain Stintzi (2013). "The transcriptional landscape of *Campylobacter jejuni* under iron replete and iron limited growth conditions". In: *PLoS One* 8.11, e79475–e79475. DOI: 10.1371/journal.pone.0079475.

Carroll, Laura M., Teresa M. Bergholz, Ian M. Hildebrandt, and Bradley P. Marks (2016). "Application of a Nonlinear Model to Transcript Levels of Upregulated Stress Response Gene *ibpA* in Stationary-Phase *Salmonella enterica* Subjected to Sublethal Heat Stress". In: *Journal of Food Protection* 79.7, pp. 1089–1096. DOI: 10.4315/0362-028X.JFP-15-377. eprint: https://doi.org/10.4315/0362-028X.JFP-15-377.

Carvalho, Daniel Cardoso, Rafael Melo Palhares, Marcela Goncalves Drummond, and Tiago Bolan Frigo (2015). "DNA Barcoding identification of commercialized seafood in South Brazil: A governmental regulatory forensic program". In: *Food Control* 50, pp. 784–788.

Carvalho, Daniel Cardoso, Rafael Melo Palhares, Marcela Goncalves Drummond, and Mario Gadanho (2017). "Food metagenomics: Next generation sequencing identifies species mixtures and mislabeling within highly processed cod products". In: *Food Control* 80, pp. 183–186.

Casey, Aidan et al. (2014). "Transcriptome analysis of *Listeria monocytogenes* exposed to biocide stress reveals a multi-system response involving cell wall synthesis, sugar uptake, and motility". In: *Front Microbiol* 5, pp. 68–68. DOI: 10.3389/fmicb.2014.00068.

Chang, Chia-Hao, Han-Yang Lin, Qiu Ren, Yeong-Shin Lin, and Kwang-Tsao Shao (2016). "DNA barcode identification of fish products in Taiwan: Government-commissioned authentication cases". In: *Food Control* 66, pp. 38–43.

Conesa, Ana et al. (2016). "A survey of best practices for RNA-seq data analysis". In: *Genome Biology* 17.1, p. 13. DOI: 10.1186/s13059-016-0881-8.

De Filippis, Francesca, Alessandro Genovese, Pasquale Ferranti, Jack A. Gilbert, and Danilo Ercolini (2016). "Metatranscriptomics reveals temperature-driven functional changes in microbiome impacting cheese maturation rate". In: *Sci Rep* 6, pp. 21871–21871. DOI: 10.1038/srep21871.

De Filippis, Francesca, Eugenio Parente, and Danilo Ercolini (2016). "Metagenomics insights into food fermentations". In: *Microb Biotechnol* 10.1, pp. 91–102. DOI: 10.1111/1751-7915.12421.

De Mattia, Fabrizio et al. (2011). "A comparative study of different DNA barcoding markers for the identification of some members of Lamiacaea". In: *Food Research International* 44.3, pp. 693–702.

Deng, Xiangyu, Zengxin Li, and Wei Zhang (2012). "Transcriptome sequencing of *Salmonella enterica* serovar Enteritidis under desiccation and starvation stress in peanut oil". In: *Food Microbiology* 30.1, pp. 311–315.

Ekblom, Robert and Jochen B. W. Wolf (2014). "A field guide to whole-genome sequencing, assembly and annotation". In: *Evolutionary Applications* 7.9, pp. 1026–1042. DOI: 10.1111/eva.12178. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/eva.12178.

Ellis, David I., Howbeer Muhamadali, David P. Allen, Christopher T. Elliott, and Royston Goodacre (2016). "A flavour of omics approaches for the detection of food fraud". In: *Current Opinion in Food Science* 10, pp. 7–15.

Ercolini, Danilo (2013). "High-throughput sequencing and metagenomics: moving forward in the culture-independent analysis of food microbial ecology". In: *Appl Environ Microbiol* 79.10, pp. 3148–3155. DOI: 10.1128/AEM.00256-13.

FDA (1998). *Bacteriological analytical manual, 8th edition, 1998 and Foodborne pathogenic microorganisms and natural toxins handbook, 1998.* Gaithersburg, MD: AOAC International.

Galia, Wessam et al. (2017). "Strand-specific transcriptomes of Enterohemorrhagic *Escherichia coli* in response to interactions with ground beef microbiota: interactions between microorganisms in raw meat". In: *BMC Genomics* 18.1, pp. 574–574. DOI: 10.1186/s12864-017-3957-2.

Galimberti, Andrea et al. (2013). "DNA barcoding as a new tool for food traceability". In: *Food Research International* 50.1, pp. 55–63.

Giordano, Francesca et al. (2017). "*De novo* yeast genome assemblies from MinION, PacBio and MiSeq platforms". In: *Scientific Reports* 7.1, p. 3935. DOI: 10.1038/s41598-017-03996-z.

Glenn, Travis C. (2011). "Field guide to next-generation DNA sequencers". In: *Molecular Ecology Resources* 11.5, pp. 759–769. DOI: 10.1111/j.1755-0998.2011.03024.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1755-0998.2011.03024.x.

Goodwin, Sara, John D. McPherson, and W. Richard McCombie (2016). "Coming of age: ten years of next-generation sequencing technologies". In: *Nature Reviews Genetics* 17, pp. 333–351.

Hellberg, Rosalee S., Brenda C. Hernandez, and Eduardo L. Hernandez (2017). "Identification of meat and poultry species in food products using DNA barcoding". In: *Food Control* 80, pp. 23–28.

Hollingsworth, Peter M., Sean W. Graham, and Damon P. Little (2011). "Choosing and Using a Plant DNA Barcode". In: *PLOS ONE* 6.5, e19254. DOI: 10.1371/journal.pone.0019254.

Hollingsworth, Peter M., De-Zhu Li, Michelle van der Bank, and Alex D. Twyford (2016). "Telling plant species apart with DNA: from barcodes to genomes". In: *Philos Trans R Soc Lond B Biol Sci* 371.1702, p. 20150338. DOI: 10.1098/rstb.2015.0338.

Hultman, Jenni, Riitta Rahkila, Javeria Ali, Juho Rousu, and K. Johanna Bjorkroth (2015). "Meat Processing Plant Microbiome and Contamination Patterns of Cold-Tolerant Bacteria Causing Food Safety and Spoilage Risks in the Manufacture of Vacuum-Packaged Cooked Sausages". In: *Applied and Environmental Microbiology* 81.20. Ed. by H. L. Drake, pp. 7088–7097. DOI: 10.1128/AEM.02228-15. eprint: https://aem.asm.org/content/81/20/7088.full.pdf.

Jackson, Brendan R. et al. (2016). "Implementation of Nationwide Real-time Whole-genome Sequencing to Enhance Listeriosis Outbreak Detection and Investigation". In: *Clinical Infectious Diseases* 63.3, pp. 380–386. DOI: 10.1093/cid/ciw242. eprint: http://oup.prod.sis.lan/cid/article-pdf/63/3/380/8039807/ciw242.pdf.

Janda, J. Michael and Sharon L. Abbott (2007). "16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls". In: *J Clin Microbiol* 45.9, pp. 2761–2764. DOI: 10.1128/JCM.01228-07.

Jarvis, Karen G. et al. (2015). "Cilantro microbiome before and after nonselective pre-enrichment for *Salmonella* using 16S rRNA and metagenomic sequencing". In: *BMC Microbiology* 15.1, p. 160. DOI: 10.1186/s12866-015-0497-2.

Jia, Kun et al. (2017). "Preliminary Transcriptome Analysis of Mature Biofilm and Planktonic Cells of *Salmonella* Enteritidis Exposure to Acid Stress". In: *Front Microbiol* 8, pp. 1861–1861. DOI: 10.3389/fmicb.2017.01861.

Kane, Dawn E. and Rosalee S. Hellberg (2016). "Identification of species in ground meat products sold on the U.S. commercial market using DNA-based methods". In: *Food Control* 59, pp. 158–163.

Kergourlay, Gilles, Bernard Taminiau, Georges Daube, and Marie-Christine Champomier Vergès (2015). "Metagenomic insights into the dynamics of mi-

crobial communities in food". In: *International Journal of Food Microbiology* 213, pp. 31–39.

Kim, Heejoong et al. (2015). "Utility of Stable Isotope and Cytochrome Oxidase I Gene Sequencing Analyses in Inferring Origin and Authentication of Hairtail Fish and Shrimp". In: *Journal of Agricultural and Food Chemistry* 63.22. PMID: 25980806, pp. 5548–5556. DOI: 10.1021/acs.jafc.5b01469. eprint: https://doi.org/10.1021/acs.jafc.5b01469.

Kovac, Jasna, Henk den Bakker, Laura M. Carroll, and Martin Wiedmann (2017). "Precision food safety: A systems approach to food safety facilitated by genomics tools". In: *TrAC Trends in Analytical Chemistry* 96.Supplement C, pp. 52–61.

Kumar, S., T. Kahlon, and S. Chaudhary (2011). "A rapid screening for adulterants in olive oil using DNA barcodes". In: *Food Chemistry* 127.3, pp. 1335–1341.

Lessard, Marie-Helene, Catherine Viel, Brian Boyle, Daniel St-Gelais, and Steve Labrie (2014). "Metatranscriptome analysis of fungal strains *Penicillium camemberti* and *Geotrichum candidum* reveal cheese matrix breakdown and potential development of sensory properties of ripened Camembert-type cheese". In: *BMC Genomics* 15, pp. 235–235. DOI: 10.1186/1471-2164-15-235.

Liao, Yu-Chieh, Shu-Hung Lin, and Hsin-Hung Lin (2015). "Completing bacterial genome assemblies: strategy and performance comparisons". In: *Scientific Reports* 5, p. 8747.

Liu, Lin et al. (2012). "Comparison of Next-Generation Sequencing Systems". In: *Journal of Biomedicine and Biotechnology* 2012. DOI: 10.1155/2012/251364.

Lusk, Tina S. et al. (2012). "Characterization of microflora in Latin-style cheeses by next-generation sequencing technology". In: *BMC Microbiol* 12, pp. 254–254. DOI: 10.1186/1471-2180-12-254.

Lynch, Susan V. and Oluf Pedersen (2016). "The Human Intestinal Microbiome in Health and Disease". In: *New England Journal of Medicine* 375.24. PMID:

27974040, pp. 2369–2379. DOI: 10.1056/NEJMra1600266. eprint: https://doi.org/10.1056/NEJMra1600266.

McClure, Ryan et al. (2013). "Computational analysis of bacterial RNA-Seq data". In: *Nucleic Acids Res* 41.14, e140–e140. DOI: 10.1093/nar/gkt444.

Monnet, Christophe et al. (2016). "Investigation of the Activity of the Microorganisms in a Reblochon-Style Cheese by Metatranscriptomic Analysis". In: *Front Microbiol* 7, pp. 536–536. DOI: 10.3389/fmicb.2016.00536.

Mudge, Jonathan M. and Jennifer Harrow (2016). "The state of play in higher eukaryote gene annotation". In: *Nat Rev Genet* 17.12, pp. 758–772. DOI: 10.1038/nrg.2016.119.

Naaum, Amanda M. et al. (2018). "Complementary molecular methods detect undeclared species in sausage products at retail markets in Canada". In: *Food Control* 84, pp. 339–344.

Noyes, Noelle R et al. (2016). "Resistome diversity in cattle and the environment decreases during beef production". In: *eLife* 5. Ed. by Ben Cooper, e13195. DOI: 10.7554/eLife.13195.

Olson, Nathan D. et al. (2015). "Best practices for evaluating single nucleotide variant calling methods for microbial genomics". In: *Front Genet* 6, pp. 235–235. DOI: 10.3389/fgene.2015.00235.

Ottesen, Andrea R. et al. (2013). "Co-enriching microflora associated with culture based methods to detect *Salmonella* from tomato phyllosphere". In: *PLoS One* 8.9, e73079. DOI: 10.1371/journal.pone.0073079.

Oulas, Anastasis et al. (2015). "Metagenomics: tools and insights for analyzing next-generation sequencing data derived from biodiversity studies". In: *Bioinform Biol Insights* 9, pp. 75–88. DOI: 10.4137/BBI.S12462.

Pardo, Miguel Angel, Elisa Jimenez, and Begona Perez-Villarreal (2016). "Misdescription incidents in seafood sector". In: *Food Control* 62, pp. 277–283.

Postollec, Florence, Helene Falentin, Sonia Pavan, Jerome Combrisson, and Daniele Sohier (2011). "Recent advances in quantitative PCR (qPCR) applications in food microbiology". In: *Food Microbiology* 28.5, pp. 848–861.

Quigley, Lisa et al. (2016). "*Thermus* and the Pink Discoloration Defect in Cheese". In: *mSystems* 1.3. Ed. by Rachel J. Dutton. DOI: 10.1128/mSystems.00023-16. eprint: https://msystems.asm.org/content/1/3/e00023-16.full.pdf.

Rasooly, Avraham and Keith E. Herold (2008). "Food microbial pathogen detection and analysis using DNA microarray technologies". In: *Foodborne Pathog Dis* 5.4, pp. 531–550. DOI: 10.1089/fpd.2008.0119.

Ratnasingham, Sujeevan and Paul D. N. Hebert (2007). "bold: The Barcode of Life Data System (http://www.barcodinglife.org)". In: *Mol Ecol Notes* 7.3, pp. 355–364. DOI: 10.1111/j.1471-8286.2007.01678.x.

Richardson, Emily J. and Mick Watson (2012). "The automatic annotation of bacterial genomes". In: *Briefings in Bioinformatics* 14.1, pp. 1–12. DOI: 10.1093/bib/bbs007. eprint: http://oup.prod.sis.lan/bib/article-pdf/14/1/1/864359/bbs007.pdf.

Ripp, Fabian et al. (2014). "All-Food-Seq (AFS): a quantifiable screen for species in biological samples by deep DNA sequencing". In: *BMC Genomics* 15.1, p. 639. DOI: 10.1186/1471-2164-15-639.

Rossi-Tamisier, Morgane, Samia Benamar, Didier Raoult, and Pierre-Edouard Fournier (2015). "Cautionary tale of using 16S rRNA gene sequence similarity values in identification of human-associated bacterial species". In: *International Journal of Systematic and Evolutionary Microbiology* 65.6, pp. 1929–1934.

Roy, Sashwati and Chandan K. Sen (2006). "cDNA microarray screening in food safety". In: *Toxicology* 221.1, pp. 128–133. DOI: 10.1016/j.tox.2005.12.025.

Sabat, A J et al. (2013). "Overview of molecular typing methods for outbreak detection and epidemiological surveillance". In: *Eurosurveillance* 18.4, 20380. DOI: https://doi.org/10.2807/ese.18.04.20380-en.

Schoch, Conrad L. et al. (2012). "Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi". In: *Proceedings of the National Academy of Sciences* 109.16, pp. 6241–6246. DOI: 10.1073/pnas.1117018109. eprint: https://www.pnas.org/content/109/16/6241.full.pdf.

Sharpton, Thomas J. (2014). "An introduction to the analysis of shotgun metagenomic data". In: *Front Plant Sci* 5, pp. 209–209. DOI: 10.3389/fpls.2014.00209.

Shokralla, Shadi et al. (2014). "Next-generation DNA barcoding: using next-generation sequencing to enhance and accelerate DNA barcode capture from single specimens". In: *Mol Ecol Resour* 14.5, pp. 892–901. DOI: 10.1111/1755-0998.12236.

Siegwald, Lea et al. (2017). "Assessment of Common and Emerging Bioinformatics Pipelines for Targeted Metagenomics". In: *PLOS ONE* 12.1, e0169563. DOI: 10.1371/journal.pone.0169563.

Stellato, Giuseppina et al. (2016). "Overlap of Spoilage-Associated Microbiota between Meat and the Meat Processing Environment in Small-Scale and Large-Scale Retail Distributions". In: *Applied and Environmental Microbiology* 82.13. Ed. by C. A. Elkins, pp. 4045–4054. DOI: 10.1128/AEM.00793-16. eprint: https://aem.asm.org/content/82/13/4045.full.pdf.

Swetha, V. P., V. A. Parvathy, T. E. Sheeja, and B. Sasikumar (2017). "Authentication of *Myristica fragrans* Houtt. using DNA barcoding". In: *Food Control* 73, pp. 1010–1015.

Tang, Silin et al. (2015). "Transcriptomic Analysis of the Adaptation of *Listeria monocytogenes* to Growth on Vacuum-Packed Cold Smoked Salmon". In: *Appl Environ Microbiol* 81.19, pp. 6812–6824. DOI: 10.1128/AEM.01752-15.

Taylor, Angela J. et al. (2015). "Characterization of Foodborne Outbreaks of *Salmonella enterica* Serovar Enteritidis with Whole-Genome Sequencing Single Nucleotide Polymorphism-Based Analysis for Surveillance and Outbreak Detection". In: *Journal of Clinical Microbiology* 53.10. Ed. by D. J. Diekema, pp. 3334–3340. DOI: 10.1128/JCM.01280-15. eprint: https://jcm.asm.org/content/53/10/3334.full.pdf.

Valdes, Alberto, Clara Ibanez, Carolina Simo, and Virginia Garcia-Canas (2013). "Recent transcriptomics advances and emerging applications in food science". In: *TrAC Trends in Analytical Chemistry* 52, pp. 142–154.

Wang, Jun and Huijue Jia (2016). "Metagenome-wide association studies: fine-mining the microbiome". In: *Nature Reviews Microbiology* 14, pp. 508–522.

Wang, Zhong, Mark Gerstein, and Michael Snyder (2009). "RNA-Seq: a revolutionary tool for transcriptomics". In: *Nat Rev Genet* 10.1, pp. 57–63. DOI: 10.1038/nrg2484.

Yandell, Mark and Daniel Ence (2012). "A beginner's guide to eukaryotic genome annotation". In: *Nature Reviews Genetics* 13, pp. 329–342.

Yang, Xiang et al. (2016). "Use of Metagenomic Shotgun Sequencing Technology To Detect Foodborne Pathogens within the Microbiome of the Beef Production Chain". In: *Appl Environ Microbiol* 82.8, pp. 2433–2443. DOI: 10.1128/AEM.00078-16.

Zhang, Feng et al. (2014). "RNA-Seq-based transcriptome analysis of aflatoxigenic *Aspergillus flavus* in response to water activity". In: *Toxins (Basel)* 6.11, pp. 3187–3207. DOI: 10.3390/toxins6113187.

CHAPTER 2

WHOLE-GENOME SEQUENCING OF DRUG-RESISTANT *SALMONELLA ENTERICA* ISOLATES FROM DAIRY CATTLE AND HUMANS IN NEW YORK AND WASHINGTON STATES REVEALS SOURCE AND GEOGRAPHIC ASSOCIATIONS [1]

[1]FROM CARROLL, LAURA M., MARTIN WIEDMANN, HENK DEN BAKKER, JULIE SILER, STEVEN WARCHOCKI, DAVID KENT, SVETLANA LYALINA, MARGARET DAVIS, WILLIAM SISCHO, THOMAS BESSER, LORIN D. WARNICK, AND RICHARD V. PEREIRA (2017). "WHOLE-GENOME SEQUENCING OF DRUG-RESISTANT *SALMONELLA ENTERICA* ISOLATES FROM DAIRY CATTLE AND HUMANS IN NEW YORK AND WASHINGTON STATES REVEALS SOURCE AND GEOGRAPHIC ASSOCIATIONS". IN: *APPLIED AND ENVIRONMENTAL MICROBIOLOGY* 83, PP. E00140-17. DOI: HTTPS://DOI.ORG/10.1128/AEM.00140-17.

2.1 Abstract

Multidrug-resistant (MDR) *Salmonella enterica* can be spread from cattle to humans through direct contact with animals shedding *Salmonella*, as well as through the food chain, making MDR *Salmonella* a serious threat to human health. The objective of this study was to use whole-genome sequencing to compare antimicrobial-resistant (AMR) *Salmonella enterica* serovars Typhimurium, Newport, and Dublin isolated from dairy cattle and humans in Washington State and New York State at the genotypic and phenotypic levels. A total of 90 isolates were selected for the study (37 *S.* Typhimurium, 32 *S.* Newport, and 21 *S.* Dublin isolates). All isolates were tested for phenotypic antibiotic resistance to 12 drugs using Kirby-Bauer disk diffusion. AMR genes were detected in the assembled genome of each isolate using nucleotide BLAST and ARG-ANNOT. Genotypic prediction of phenotypic resistance resulted in a mean sensitivity of 97.2% and specificity of 85.2%. Sulfamethoxazole-trimethoprim resistance was observed only in human isolates ($P < 0.05$), while resistance to quinolones and fluoroquinolones was observed only in 6 *S.* Typhimurium isolates from humans in Washington State. *S.* Newport isolates showed a high degree of AMR profile similarity, regardless of source. *S.* Dublin isolates from New York State differed from those from Washington State based on the presence/absence of plasmid replicons, as well as phenotypic AMR susceptibility/nonsusceptibility ($P < 0.05$). The results of this study suggest that distinct factors may contribute to the emergence and dispersal of AMR *S. enterica* in humans and farm animals in different regions.

IMPORTANCE: The use of antibiotics in food-producing animals has been hypothesized to select for AMR *Salmonella enterica* and associated AMR deter-

minants, which can be transferred to humans through different routes. Previous studies have sought to assess the degree to which AMR livestock- and human-associated *Salmonella* strains overlap, as well as the spatial distribution of *Salmonella*'s associated AMR determinants, but have often been limited by the degree of resolution at which isolates can be compared. Here, a comparative genomics study of livestock- and human-associated *Salmonella* strains from different regions of the United States shows that while many AMR genes and phenotypes were confined to human isolates, overlaps between the resistomes of bovine and human-associated *Salmonella* isolates were observed on numerous occasions, particularly for *S.* Newport. We have also shown that whole-genome sequencing can be used to reliably predict phenotypic resistance across *Salmonella* isolated from bovine sources.

2.2 Introduction

Salmonella enterica is estimated to cause approximately 1.2 million illnesses and 450 deaths each year in the United States alone (Scallan et al. 2011). While most individuals recover without medical intervention, severe infections require hospitalization and treatment with antimicrobials (Scallan et al. 2011). An even greater challenge is posed when those infections are caused by antimicrobial-resistant (AMR) organisms. The Centers for Disease Control and Prevention (CDC) estimates that 100,000 infections due to AMR non-typhoidal *Salmonella* occur in the United States annually and has designated AMR in non-typhoidal *Salmonella* as a serious threat to public health (CDC 2013). More specifically, the World Health Organization (WHO) has listed fluoroquinolone-resistant non-typhoidal *Salmonella* as a global health concern (WHO 2014).

Both the CDC and WHO have called for improved monitoring of AMR along the food chain, particularly in food-producing animals (CDC 2013; WHO 2014). Due to concerns about the misuse of antimicrobials in farm animals, the farm is often viewed as a reservoir in which AMR can be acquired by bacteria that are then transmitted from animals to humans (Van Boeckel et al. 2015; Silbergeld, Graham, and Price 2008). In this context, *S. enterica* becomes particularly relevant, as it can be transmitted between animal and human populations (Hendriksen et al. 2004; Fey et al. 2000; Hoelzer, Moreno Switt, and Wiedmann 2011), as well as through food (White et al. 2001; Cody et al. 1999; Hald et al. 2016).

A number of studies have sought to assess the extent to which AMR is acquired by bacteria in livestock environments and subsequently transmitted to humans, and many have arrived at different conclusions (Johnson et al. 2007; Price et al. 2012; A. E. Mather et al. 2013; Alison E. Mather et al. 2012). Often, the degree of resolution at which isolates can be compared is a limiting factor in determining the origin of a particular bacterial isolate and its AMR profile. Methods such as multilocus sequence typing (MLST), serotyping, and pulsed-field gel electrophoresis (PFGE) may not offer enough discriminatory power to detect differences between isolates from different sources or locations (Kwong et al. 2016; Holmes et al. 2015; Taylor et al. 2015), while phenotypic testing of AMR may not distinguish between AMR mechanisms in different isolates (A. E. Mather et al. 2013).

The extent to which *Salmonella* and AMR genes associated with it are transmitted between animal and human sources remains unclear. The objective of this study was to use whole-genome sequencing (WGS) to compare AMR *Salmonella enterica* isolates previously serotyped as Typhimurium, Newport, or

Dublin isolated from dairy cattle and humans in Washington State and New York State at the genotypic and phenotypic levels. In addition, correlations between AMR genotype and AMR phenotype were assessed. It was hypothesized that sources and geographic differences between *Salmonella* isolates could be elucidated at greater resolution through the implementation of WGS.

2.3 Materials and Methods

2.3.1 Isolate selection

A total of 93 *Salmonella* isolates were initially selected for the study. Bovine isolates originated from the Washington Animal Disease Diagnostic Laboratory (WADDL), the Washington State Zoonotic Research Unit, the Cornell Animal Health Diagnostic Center (Ithaca, NY), and *Salmonella* strains isolated from dairy cattle during previous research sampling at dairy farms. Isolates from human clinical specimens were obtained from the Washington State Department of Health Public Health Laboratory and from the New York State Department of Health Laboratory. Isolates were selected to (i) represent isolation dates between 2008 and 2012; (ii) represent one of the three serotypes of interest (Typhimurium, Newport, and Dublin, as determined using traditional serotyping; these serotypes were selected for their association with humans and cattle); and (iii) represent isolates that had previously been tested for phenotypic resistance to antimicrobials and were found to be resistant to at least one antimicrobial. Bovine isolates originated from fecal samples, independent of whether the host presented clinical signs of salmonellosis or not, while human isolates were from

stool samples of patients presenting clinical signs of salmonellosis. Among the isolates that met these criteria, "redundant" isolates were filtered out (those known to come from the same animal/farm/farm visit), and selected isolates were chosen to represent approximately equal numbers of human and bovine isolates evenly distributed between New York State and Washington State. To ensure consistency between phenotypic testing methods, all of the isolates selected for this study were re-tested for phenotypic resistance using a single AMR testing method and a panel of antimicrobial drugs (see "Phenotypic AMR testing" below).

Following WGS (see "Whole-genome sequencing" below), seven isolates were found to belong to species/serotypes different from those to which they were initially assigned. One isolate that had been initially classified as *S. enterica* serotype Newport was found to belong to the genus *Citrobacter*. In addition, *in silico* multilocus sequence typing (MLST) and *in silico* serotyping using WGS data from the isolates (see "*In silico* serotyping and MLST" below) revealed that two of the isolates that had been classified as serotypes Typhimurium and Newport using traditional serotyping methods actually belonged to serotypes Give and Montevideo, respectively. These two isolates, as well as the *Citrobacter* isolate, were excluded from the study. Four isolates that were classified using traditional serotyping as Newport, Typhimurium, Typhimurium, and Dublin were reclassified as Dublin, Newport, Dublin, and Newport, respectively, and remained in the study under the new serotype classifications. A total of 90 isolates (37 *S.* Typhimurium, 32 *S.* Newport, and 21 *S.* Dublin isolates; see Table S1 in the supplemental material for details) were used in all subsequent analyses.

2.3.2 Phenotypic AMR testing

The antimicrobial susceptibility of each *Salmonella* isolate was tested using a modified National Antimicrobial Resistance Monitoring System (NARMS) panel of 12 antimicrobial drugs. Susceptibility testing was performed using a Kirby-Bauer disk diffusion agar assay in accordance with the guidelines published by the Clinical and Laboratory Standards Institute (CLSI) and a methodology previously described (CLSI 2012; CLSI 2013). Internal quality control was performed by the inclusion of *E. coli* ATCC 25922, which had previously been determined to be pan-susceptible, as well as an *E. coli* isolate that had been previously characterized as positive for the $bla_{CMY\text{-}2}$ gene and resistant to nine of the antimicrobial agents tested. All isolates were tested using the following panel: ampicillin (AMP) at 10 μg, amoxicillin-clavulanic acid (AMC) at 20 and 10 μg, respectively, cefoxitin (FOX) at 30 μg, ceftiofur (TIO) at 30 μg, ceftriaxone (CRO) at 30 μg, chloramphenicol (CHL) at 30 μg, ciprofloxacin (CIP) at 5 μg, nalidixic acid (NAL) at 30 μg, streptomycin (STR) at 10 μg, tetracycline (TET) at 30 μg, sulfisoxazole (SX) at 250 μg, and trimethoprim-sulfamethoxazole (SXT) at 23.75 and 1.25 μg, respectively. Results of the disk diffusion test for the internal quality control strains were within the anticipated standards. Isolates were categorized as susceptible, intermediate, or resistant (SIR) by measuring the inhibition zone and using interpretive criteria and breakpoints established by the CLSI guidelines for each antimicrobial (CLSI 2012).

2.3.3 Whole-genome sequencing

Isolates were plated on brain heart infusion (BHI) agar (Becton, Dickinson and Company, Franklin Lakes, NJ), grown for 24 h, and inoculated into 1.0 ml BHI broth in a Nunc U96 PP 2-ml DeepWell Natural plate (Fisher Scientific, Pittsburgh, PA). Following overnight incubation at 37°C, cells were pelleted by centrifugation at 3,320 relative centrifugal force (RCF) for 15 min. DNA extraction for the majority of isolates was performed with the DNeasy 96 blood and tissue kit (Qiagen, Valencia, CA) according to the manufacturer's specifications for high-throughput applications. DNA extraction for a smaller group of isolates was performed using the QIAamp DNA minikit (Qiagen, Valencia, CA) according to the manufacturer's protocol for bacteria. DNA was eluted in 50 μl Tris-HCl at pH 8.0 and stored at 4°C prior to sequencing. Following an initial spectrophotometry step to determine the optical density at 260 nm (OD_{260})/OD_{280} measurements, the genomic DNA from each isolate was quantified using a fluorescent nucleic acid dye (Picogreen; Invitrogen, Paisley, UK) and diluted to 200 pg/μl. Sequencing libraries were prepared using the Nextera XT DNA sample preparation kit and the associated Nextera XT Index kit with 96 indices (Illumina, Inc., San Diego, CA) according to the manufacturer's instructions. Pooled samples were sequenced with 2 lanes of an Illumina HiSeq 2500 rapid run with 2 x 100-bp paired-end sequencing.

2.3.4 Initial data processing and genome assembly

Illumina sequencing adapters and low-quality bases were trimmed using Trimmomatic version 0.32 for Nextera paired-end reads (Bolger, Lohse, and Usadel

2014). FastQC version 0.11.2 was used to confirm that all adapter sequences had been removed and that the read quality was appropriate (Andrews 2014). Genomes were assembled *de novo* using SPAdes version 3.0.0, as SPAdes has been shown to produce few misassemblies and yield contigs with high N50 values when assembling bacterial genomes *de novo* from Illumina short reads (Bankevich et al. 2012). Genome coverage was determined using BBMap version 35.49 (Bushnell 2015) and samtools version 0.1.19-96b5f2294a (H. Li et al. 2009).

2.3.5 *In silico* serotyping and MLST

To assess the results of traditional serotyping, *in silico* serotyping was performed using SeqSero and the assembled genome for each isolate (Zhang et al. 2015). In addition, MLST was performed using the Short Read Sequence Typer 2 version 0.1.5 (SRST2) and the trimmed Illumina paired-end reads (Inouye et al. 2014). Sequence types were associated with serotypes using the University of Warwick's MLST database for *Salmonella* (http://mlst.warwick.ac.uk).

2.3.6 *In silico* AMR gene detection

AMR genes were detected in all 90 assembled genomes using nucleotide BLAST (blastn) version 2.4.0 (Camacho et al. 2009) and the formatted ARG-ANNOT database included with SRST2 (Inouye et al. 2014; Gupta et al. 2014). To prevent overlapping hits due to the presence of multiple alleles of the same gene in the database, one gene was selected from each SRST2-ARG-ANNOT gene group

and used to build a reduced database (Inouye et al. 2014). Genes that were detected using blastn and belonged to a particular gene group were categorized as being present in a genome if they were detected at 50% coverage and 75% nucleotide identity.

2.3.7 Initial phylogenetic tree construction and reference genome selection

The closed chromosomal sequences of *S.* Typhimurium strain LT2 (RefSeq NC_003197.1), *S.* Newport strain SL254 (GenBank accession no. CP001113), and *S.* Dublin strain CT_02021853 (RefSeq NC_011205.1) were chosen as candidate reference sequences for reference-based SNP calling. To obtain an initial phylogeny of all isolates and determine if these candidate reference sequences clustered appropriately with the genomes of the isolates used in this study, a phylogenetic tree was constructed using the assembled genomes of all 90 isolates and the three candidate reference genomes using kSNP version 2.1.2 (Gardner and Hall 2013). Kchooser was used to determine an optimum *k*-mer size of 19 (Gardner and Hall 2013). This core SNP phylogeny based on the genomes of all 90 isolates used in the study, as well as three closed reference genomes from GenBank, clustered isolates into three distinct clades (see Fig. S1 in the supplemental material). As a result, all subsequent analyses were performed within each serotype clade to maximize resolution.

2.3.8 Reference-based variant calling

Variant calling was performed within each of the three serotypes using the Cortex variant caller (cortex_var) (Iqbal et al. 2012). For *S.* Typhimurium isolates, *S.* Typhimurium strain LT2 was used as a reference genome. For *S.* Newport isolates, *S.* Newport strain SL254 was used as a reference, as all of the Newport isolates in this study were predicted to have the same sequence type (ST45) using SRST2 (Inouye et al. 2014). For *S.* Dublin isolates, strain CT_02021853, which was used as a candidate reference in the initial phylogenetic tree, clustered relatively far from the closely related *S.* Dublin isolates used in this study. In order to obtain better resolution, variant calling was performed a second time using the contigs of isolate BOV_DUBN_WA_10_R9_3233 as a reference, as its assembly had the highest coverage of all of the *S.* Dublin isolates used in the study. An additional 11 SNPs were found using isolate BOV_DUBN_WA_10_R9_3233 as a reference; these SNPs were included in subsequent analyses. SNPs were filtered from other variants using Plink/Seq version 0.10 (PLINK/Seq 2014), and recombination events were filtered out using Gubbins version 1.4.2 (Croucher et al. 2015). Within each serotype, only SNPs at positions present in all genomes were used. MEGA6 was used to identify the best nucleotide substitution models for SNPs within each serotype (Tamura et al. 2013). For *S.* Typhimurium, the general time-reversible (GTR) model was selected as the best model (Tavare n.d.), while the Kimura 2-parameter model (Kimura 1980) was selected for both *S.* Newport and *S.* Dublin.

For each serotype, BEAST version 1.8.2 (Alexei J. Drummond et al. 2012) was used to construct rooted phylogenetic trees. An ascertainment bias correction was applied to account for the use of solely variant sites (Rambaut 2013).

The best nucleotide substitution model, as determined by MEGA6, was used for each serotype, and base frequencies were estimated. Temporal signals, which were assessed using Path-O-Gen version 1.4 (now TempEst) (Rambaut et al. 2016), were not strong enough to estimate evolutionary rates using sampling dates ($R < 0.10$). As a result, the clock rate was set to 1.0 and tip dates were not used. For each serotype, combinations of either a strict or lognormal relaxed molecular clock (A. J. Drummond, Ho, et al. 2006) and either a coalescent constant size or Bayesian skyline population (A. J. Drummond, Rambaut, et al. 2005) were tested. Trees were constructed using chain lengths of 100 million generations, with sampling every 10,000 generations. Path sampling analyses (Baele, Lemey, et al. 2012; Baele, W. L. S. Li, et al. 2013) were performed using 100 steps of 1 million generations, sampling every 1,000 generations. Bayes factors were calculated to determine which combination of molecular clock and population models best modeled each serotype. For *S.* Typhimurium and *S.* Newport, the best model used a relaxed molecular clock with a constant coalescent population model. For *S.* Dublin, the best model used a strict molecular clock with a constant coalescent population.

2.3.9 Plasmid replicon detection

Plasmid replicons were detected in all whole-genome sequences using PlasmidFinder version 1.3 (Carattoli et al. 2014). An identity cutoff of 80% was used. PlasmidFinder was also used to confirm that plasmid replicons could not be detected in the chromosomal sequences of *S.* Typhimurium LT2, *S.* Newport SL254, and *S.* Dublin CT_02021853.

2.3.10 Statistical analyses

Matrices were created using (i) the sequences of all AMR genes detected using blastn, (ii) phenotypic antimicrobial resistance/susceptibility, and (iii) the presence/absence of plasmid replicons detected using PlasmidFinder. For the phenotypic resistance matrix, isolates showing resistance or intermediate resistance to a particular antimicrobial, using NARMS breakpoints, were treated as resistant and given a value of 1, while susceptible isolates were given a value of 0. Fisher's exact tests were conducted to test whether a given AMR gene, AMR phenotype, or plasmid replicon was statistically associated with a particular source and/or geographic location using the fisher.test function in R version 3.3.0 (R Core Team 2016). When performing Fisher's exact tests for each serotype category with n isolates, gene groups, AMR phenotypes, and plasmid replicons present in fewer than 3 and more than $n - 3$ isolates were not tested. A Holm-Bonferroni correction was applied to each test to correct for multiple comparisons (Holm 1979). Additionally, Fisher's exact tests were used to test if any AMR gene groups were statistically associated with any plasmid replicons. Plasmid replicons present in fewer than 5 and more than $n - 5$ isolates were not tested, and a Bonferroni correction was applied to correct for multiple comparisons. Analysis of similarity (ANOSIM) (Clarke 1993) using the anosim function in the vegan package (Oksanen et al. 2017) in R was used to determine if the average ranks of within-serotype, within-source, and within-geographic-group distances were greater than or equal to the average ranks of between-group distances using AMR gene sequences, phenotypic resistance to a particular antimicrobial, and/or plasmid replicon presence/absence data (Anderson and Walsh 2013). For ANOSIM simulations using AMR gene sequences, 5 runs of 10,000 permutations using unweighted unifrac dis-

tances (Lozupone and Knight 2005) were conducted. For all ANOSIM simulations using phenotypic resistance/susceptibility and plasmid replicon presence/absence matrices, 5 runs of 10,000 permutations using Raup-Crick dissimilarities (Chase et al. 2011) were conducted. PERMANOVA (Anderson 2001) was performed to test whether the centroids of serotype, source, and geographic groups were equivalent for all groups (Anderson and Walsh 2013) based on AMR gene sequences, phenotypic resistance to a particular antimicrobial, and/or plasmid replicon presence/absence using the adonis function in R's vegan package (Oksanen et al. 2017). Three runs of 10,000 permutations using unweighted unifrac distances were used to obtain mean PERMANOVA test statistics (F) and P values for AMR gene sequences, while three runs of 100,000 permutations and Raup-Crick distances were used for phenotypic resistance/susceptibility and plasmid replicon presence/absence data. The metaMDS function in the vegan package was used to perform nonmetric multidimensional scaling (NMDS) (Kruskal 1964a; Kruskal 1964b) using monoMDS (Oksanen et al. 2017), a maximum of 10,000 random starts, and an appropriate distance metric (unweighted unifrac distances for AMR gene sequence data and Raup-Crick dissimilarities for phenotypic resistance/susceptibility and plasmid replicon presence/absence data). Interactive NMDS plots can be found at https://github.com/lmc297/2017_AEM_Figure_S2.

Descriptive analyses of the susceptible/intermediate/resistant (SIR) distribution of *Salmonella* isolates by antimicrobial drug and distribution of AMR phenotypes and genes were performed using PROC FREQ in SAS (SAS Institute Inc., USA). To evaluate the effect of presence or absence of resistance genes on the mean zone diameter (in centimeters) of the Kirby-Bauer disk diffusion test, multivariable mixed logistic regression models were fitted to the data us-

ing the Glimmix procedure of SAS. The independent variables (i) isolate source (bovine or human), (ii) isolation location (New York State or Washington State), and (iii) serotype were included in all models.

2.3.11 Accession number(s) and supplemental material

Paired-end reads for the 90 isolates used in this study have been deposited in the National Center for Biotechnology Information's (NCBI) Sequence Read Archive (SRA) under study accession number SRP068320. Supplemental material for this article may be found at https://doi.org/10.1128/AEM.00140-17.

2.4 Results

2.4.1 Overall distribution of SNPs, AMR genes, AMR phenotypes, and plasmid replicons

Of the three serotypes studied, *S.* Typhimurium displayed the highest degree of phylogenetic diversity. Variant calling revealed a total number of 2,976 variants in the *S.* Typhimurium isolates, with 2,723 of those variants called as single nucleotide polymorphism (SNPs). In *S.* Newport, only 327 variants were called, 263 of which were SNPs. The fewest number of variants occurred in *S.* Dublin, with 183 variants, 131 of which were SNPs.

AMR genes belonging to 42 different groups were detected in the 90 genomes (see Table S2 in the supplemental material). The most common genes

belonged to groups associated with resistance to penicillins (penicillin binding protein [*PBP*] gene), aminoglycosides [*aac(6)-Iaa*, *strA*, and *strB*], phenicols (*floR*), tetracyclines [*tet(A)* and *tet(R)*], cephalosporins (*CMY*), and sulfonamides (*sul2*) (Table 2.1). At the phenotypic level, all isolates displayed resistance or intermediate resistance to between 1 and 11 antimicrobials. The most common antimicrobial to which isolates were resistant was ampicillin (AMP), as 88 of 90 isolates were AMP resistant (Table 2.1). In addition, a total of 20 different plasmid replicons were detected in the genomes of the 90 isolates used in the study. The three most common replicons (ColRNAI, ColpVC, and IncA/C2) were each detected in over one-half of all isolates (Table 2.1). Several significant ($P < 0.001$) associations between plasmid replicons and AMR gene groups were observed, including the IncA/C2 replicon and gene groups *CMY*, *floR*, *strA-strB*, *sul2*, and *tet(A)-tet(R)* (see Table S3 in the supplemental material). These genes had previously been found on an IncA/C2 plasmid isolated from *S.* Newport (Fricke et al. 2009).

Serotypes were found to differ with regard to AMR gene sequences, phenotypic resistance/susceptibility, and the presence/absence of plasmid replicons when using analysis of similarity (ANOSIM) and/or permutational multivariate analysis of variance (PERMANOVA; $P < 0.001$ after a Holm-Bonferroni correction) (Table 2.2). Of the three serotypes studied, S. Typhimurium showed the widest range of AMR gene profiles, phenotypic AMR profiles, and plasmid replicon presence/absence profiles (Figure 2.1).

Table 2.1: Ranking of the five most common antimicrobial resistance (AMR) gene groups, phenotypic AMR profiles, and plasmid replicons for all serotypes, *S.* Typhimurium, *S.* Newport, and *S.* Dublin[a]

Rank[b]	***All isolates*** *(n = 90)*	***S. Typhimurium*** *(n = 37)*	***S. Newport*** *(n = 32)*	***S. Dublin*** *(n = 21)*
AMR gene groups				
1	*aac(6)-Iaa*, *PBP* gene (90)	*aac(6)-Iaa*, *PBP* gene (37)	*aac(6)-Iaa*, *CMY*, *PBP* gene, *strA*, *strB*, *sul2*, *tet(A)*, *tet(R)* (32)	*aac(6′)-Iaa*, *CMY*, *PBP* gene, *sul2* (21)
2	*floR* (72)	*aadA* (25)	*floR* (30)	*strA*, *strB*, *tet(A)*, *tet(R)* (20)
3	*CMY*, *tet(A)*, *tet(R)* (68)	*floR* (23)	*aph(3″)-Ia* (22)	*floR* (19)
4	*sul2* (67)	*sul1* (21)	*aadA*, *dfrA*, *sul1* (3)	*aph(3″)-Ia* (18)
5	*strA*, *strB* (64)	*aph(3″)-Ia* (20)		$bla_{TEM\text{-}1D}$ (15)
Phenotypic AMR profile				
1	AMP (88)	AMP (35)	AMC; AMP; CRO; FOX; STR; SX; TIO; TET (32)	AMP; CRO; TIO (21)
2	TET (82)	TET (31)	CHL (30)	AMC; FOX; SX (20)
3	AMC; SX (81)	STR (30)	SXT (3)	CHL; TET (19)
4	CHL; STR (72)	AMC; SX (29)		STR (10)
5	CRO; TIO (71)	CHL (23)		SXT (1)
Plasmid replicons				
1	ColRNAI (77)	ColRNAI (27)	ColRNAI; IncA/C2 (32)	IncX1 (21)
2	ColpVC (63)	IncFII(S) (25)	ColpVC (26)	IncA/C2 (20)
3	IncA/C2 (60)	ColpVC (20)	IncI1 (2)	ColRNAI (18)
4	IncFII(S) (36)	IncFIB(S) (17)	Col(BS512) (1)	ColpVC (17)
5	IncX1 (22)	IncI1 (10)		IncFII(S) (11)

[a]Numbers in parentheses indicate the number of isolates (i) carrying genes classified into a given AMR gene group, (ii) resistant to a given antimicrobial, or (iii) carrying a given plasmid replicon.
[b]Rank is based on the frequency of (i) AMR gene group presence, (ii) phenotypic resistance, and (iii) plasmid replicon presence.

2.4.2 *In silico* AMR gene detection is correlated with phenotypic AMR patterns.

Genotypic and phenotypic AMR data were used to evaluate the ability of genotypic data to predict phenotypic resistance (Figure 2.2). Ciprofloxacin (CIP) was not included in these analyses due to the rarity of resistant isolates in this data set (1 of the 90 isolates). Based on the 11 remaining antimicrobials, genotypic prediction of phenotypic resistance resulted in a mean sensitivity of 97.2% and specificity of 85.2% (Table 2.3). Genotypic prediction of phenotypic resistance to AMP, cefoxitin (FOX), chloramphenicol (CHL), streptomycin (STR), sulfisoxazole (SX), and tetracycline (TET) had a sensitivity of 100%, while the prediction of phenotypic resistance to AMP, ceftiofur (TIO), ceftriaxone (CRO), nalidixic acid (NAL), and trimethoprim-sulfamethoxazole (SXT) had a specificity of 100% (Table 2.3). With the exception of NAL, genotypic prediction of phenotypic re-

Table 2.2: ANOSIM and PERMANOVA statistics and their respective mean *P* values[a]

		ANOSIM		PERMANOVA	
Serotype(s)	*Grouping factor/response*[b]	*R statistic*	*Mean uncorrected P value*	*F statistic*	*Mean uncorrected P value*
Antimicrobial resistance gene sequences					
All	**Serotype**	0.234[c]	< 0.001[c]	15.598[d]	< 0.001[d]
Typhimurium	Source	0.079	0.040	2.937	0.020
Typhimurium	Location	0.045	0.105	2.093	0.074
Newport	Source	0.034	0.169	3.405	0.004
Newport	**Location**	0.241[c]	0.002[c]	3.185	0.008
Dublin	Source	0.041	0.188	1.578	0.231
Dublin	Location	0.145	0.064	5.366	0.004
Phenotypic antimicrobial resistance/susceptibility profiles					
All	**Serotype**	0.200[c]	< 0.001[c]	1.037	0.433
Typhimurium	Source	0.122	0.015	6.796	0.012
Typhimurium	Location	−0.003	0.417	0.181	0.727
Newport	Source	−0.030	1.000	1.739	0.053
Newport	Location	0.103	0.072	1.699	0.074
Dublin	Source	0.089	0.053	1.060	0.477
Dublin	**Location**	0.481[c]	< 0.001[c]	4.717[d]	< 0.001[d]
Plasmid replicon presence/absence profiles					
All	**Serotype**	0.350[c]	< 0.001[c]	21.800[d]	< 0.001[d]
Typhimurium	Source	0.025	0.201	−0.299	0.853
Typhimurium	Location	0.107	0.009	6.077	0.011
Newport	Source	−0.030	0.934	2.118	0.042
Newport	Location	0.098	0.074	1.572	0.105
Dublin	Source	0.040	0.146	1.521	0.116
Dublin	**Location**	0.408[c]	< 0.001[c]	4.466[d]	< 0.001[d]

[a]Rows in boldface indicate that at least one test was significant ($P < 0.05$) after a Holm-Bonferroni correction was applied.
[b]Grouping factors used were serotype (only for "All isolates"), source (bovine or human), and location (New York or Washington State).
[c]Significant ANOSIM test ($P < 0.05$) after a Holm-Bonferroni correction was applied.
[d]Significant PERMANOVA test ($P < 0.05$) after a Holm-Bonferroni correction was applied.

sistance resulted in sensitivities greater than 90% for all drugs (Table 2.3). For all antimicrobials other than AMC, STR, SX, and TET, genotypic prediction of phenotypic resistance had specificity above 90% (Table 2.3). Consistent with these findings, significant differences in resistance (determined by the mean zone diameters from the Kirby-Bauer disk diffusion assays) were observed between isolates carrying at least one AMR gene conferring resistance to a given antimicrobial and those isolates that did not carry said AMR gene ($P < 0.05$ after a Holm-Bonferroni correction) (Table 2.4).

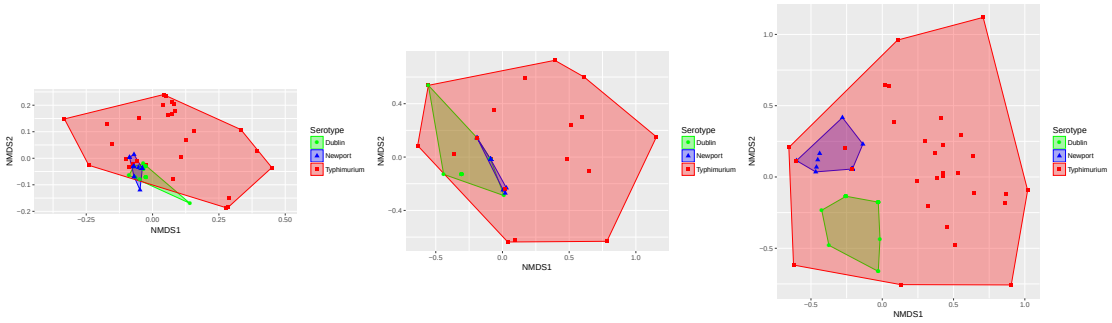


Figure 2.1: Nonmetric multidimensional scaling (NMDS) plots for all isolates based on antimicrobial resistance (AMR) gene sequences (A), phenotypic antimicrobial resistance/susceptibility profiles (B), and presence/absence of plasmid replicons (C). Points represent isolates, while shaded regions and convex hulls correspond to isolate serotypes. For an interactive plot of these data, as well as interactive NMDS plots for individual serotypes, visit https://github.com/lmc297/2017_AEM_Figure_S2.

Table 2.3: Sensitivity and specificity of genotype predictions of AMR phenotype for all 90 *Salmonella* isolates in the study.

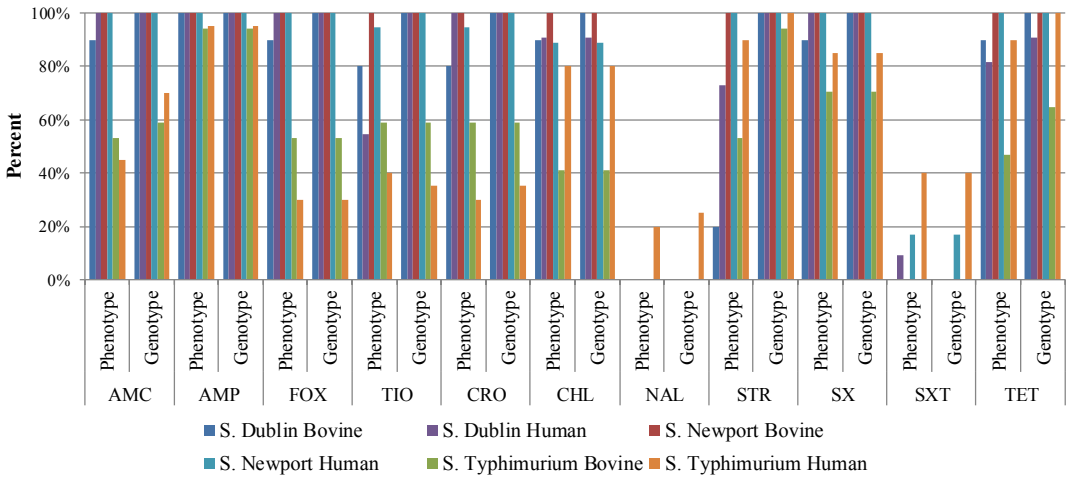
Antimicrobial[a]	**Phenotype: resistant** (*n*)[b]		**Phenotype: susceptible** (*n*)		*Sensitivity (%)*	*Specificity (%)*
	Genotype: resistant	*Genotype: susceptible*	*Genotype: resistant*	*Genotype: susceptible*		
AMC	71	2	6	11	97.3	64.7
AMP	88	0	0	2	100.0	100.0
FOX	67	0	1	22	100.0	95.7
TIO	70	1	0	19	98.6	100.0
CRO	70	1	0	19	98.6	100.0
CHL	72	0	1	17	100.0	94.4
NAL	5	1	0	84	83.3	100.0
STR	72	0	17	1	100.0	5.6
SX	81	0	1	8	100.0	88.9
SXT	11	1	0	78	91.7	100.0
TET	82	0	1	7	100.0	87.5
Overall					97.2	85.2

[a] AMC, amoxicillin-clavulanic acid; AMP, ampicillin; FOX, cefoxitin; TIO, ceftiofur; CRO, ceftriaxone; CHL, chloramphenicol; NAL, nalidixic acid; STR, streptomycin; SX, sulfisoxazole; SXT, sulfamethoxazole/trimethoprim; TET, tetracycline

[b] Isolates that showed intermediate resistance to an antimicrobial are categorized as resistant.

2.4.3 *S.* Typhimurium phylogeny, AMR genes, AMR phenotypes, and plasmid replicons

A BEAST phylogeny of the 37 *S.* Typhimurium genomes separated these isolates into two major clades (Figure 2.3; posterior probability,



AMC, amoxicillin/clavulanic acid; AMP, ampicillin; FOX, cefoxitin; TIO, Ceftiofur; CRO, ceftriaxone; CHL, chloramphenicol; STR, streptomycin; SX, sulfisoxazole; SXT, sulfamethoxazole / trimethoprim; TET, tetracycline.

Figure 2.2: Frequency of different phenotypic and genotypic resistance determinants for each serotype-source group (e.g., *Salmonella* Dublin isolates obtained from humans [*S.* Dublin Human]). Genotypic resistance was determined using nucleotide BLAST (blastn) and the ARG-ANNOT database; isolates were classified as having a resistant genotype if the AMR gene was detected by BLAST with a minimum coverage of 50% and a minimum sequence identity of 75%. Phenotypic resistance was tested using Kirby-Bauer disk diffusion. Percentages were calculated using the ratio of resistant isolates to total isolates in each serotype-source group (n = 17 for *S.* Typhimurium Bovine, n = 20 for *S.* Typhimurium Human, n = 14 for *S.* Newport Bovine, n = 18 for *S.* Newport Human, n = 10 for *S.* Dublin Bovine, and n = 11 for *S.* Dublin Human). Nalidixic acid (NAL)- and sulfamethoxazole-trimethoprim (SXT)-resistant isolates (6 and 12 of the 90 isolates, respectively) each had one isolate for which genotypic resistance did not correlate with phenotypic resistance.

1). One of these clades contained human isolates exclusively (n = 8), while the other major clade included 12 human and 17 bovine isolates (Figure 2.3). Three isolates within this "mixed source" clade were particularly similar based on their AMR gene sequences: isolates BOV_TYPH_WA_09_R9_3247 (isolated from a dairy cow in Washington State in 2009), HUM_TYPH_WA_09_R9_3271 (isolated from a human in Washington State in 2009), and HUM_TYPH_NY_12_R9_0437 (isolated from a human in New York State in 2012) appeared to have highly similar AMR gene profiles (see Fig-

Table 2.4: Comparison of mean zone diameters between (i) *Salmonella* isolates with at least one AMR gene (ARG) that has been known to confer resistance to a particular antimicrobial and (ii) isolates with no genes known to confer resistance to that antimicrobial.[a]

	95% CI of MZD[a] (cm)	
Antimicrobial	***ARG absent***	***ARG present***
Aminopenicillins		
Ampicillin	25.4-25.6	0.0-0.02
Amoxicillin-clavulanic acid	13.9-18.7	9.2-11.0
Chloramphenicol	24.4-27.6	0.02-1.45
Cephalosporins		
Ceftiofur	25.5-29.5	12.7-14.5
Ceftriaxone	29.7-34.5	13.4-15.5
Cefoxitin	23.2-27.5	8.4-10.2
Streptomycin	13.9-21.1	3.1-5.3
Sulfonamides		
Sulfisoxazole	22.4-26.2	0.0-0.9
Sulfamethoxazole-trimethoprim	23.8-25.8	0-3.3
Tetracycline	19.0-26.5	2.0-4.2

[a]MZD, mean zone diameter; CI, confidence interval. All *P* values were < 0.0001.

ure S2 posted at https://github.com/lmc297/2017_AEM_Figure_S2). All AMR genes in these three isolates matched with 100% sequence identity except for *tet(RG)*; HUM_TYPH_WA_09_R9_3271 *tet(RG)* differed from the other two isolates at nucleotide position 73.

Overall, 41 of the 42 AMR gene groups identified in the 90 isolates in this study were detected in *S.* Typhimurium (all except *aadB*; Figure 2.3). The 37 *S.* Typhimurium isolates were distributed into 24 different genotypic MDR profiles, the most common of which was *aac(6)-Iaa floR sul1 tet(RG) tet(G) bla$_{CARB}$ aadA PBP* gene, which was found in 11% of *S.* Typhimurium genomes. In addition, between 2 and 7 unique plasmid replicons were detected per genome (Figure 2.3). When ANOSIM and PERMANOVA were applied as metrics to assess clustering based on either AMR gene sequences or plasmid replicon presence/absence, there were no significant differences between bovine and human isolate clusters or between New York and Washington State clusters ($P > 0.05$ after a Holm-Bonferroni correction) (Table 2.2). While neither ANOSIM nor PER-

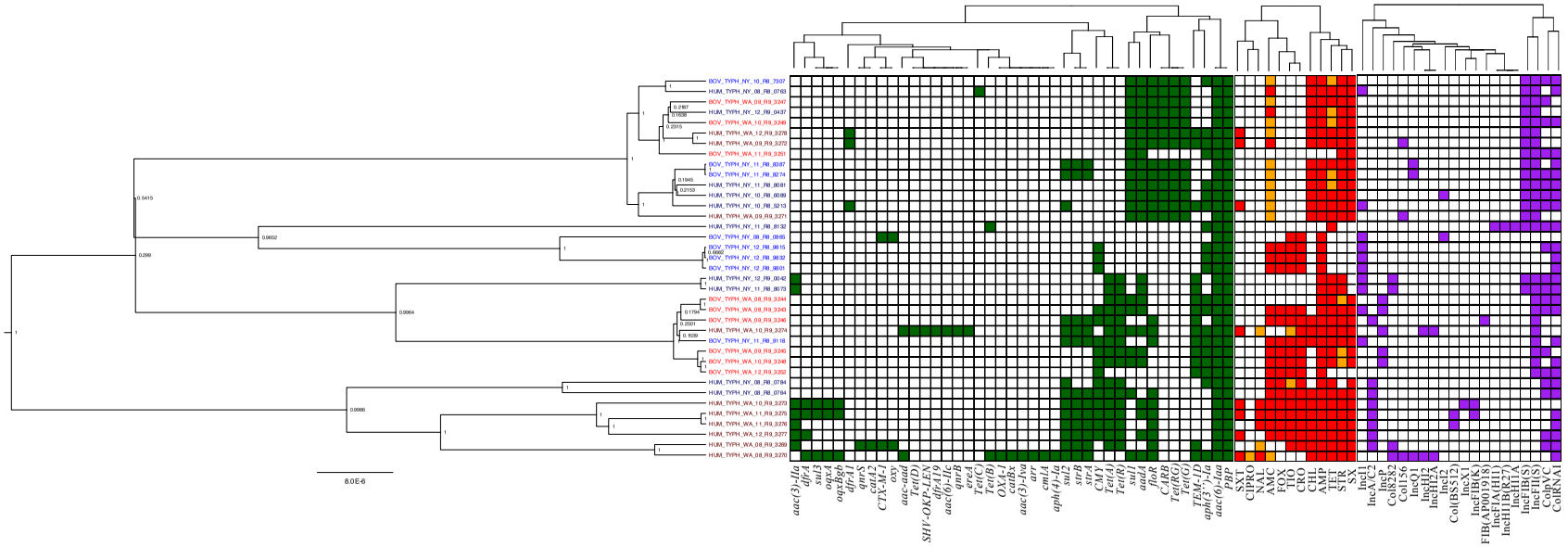


Figure 2.3: Phylogenetic tree of *S.* Typhimurium isolates constructed using BEAST. Gene groups for AMR genes detected in each genome sequence at more than 50% coverage and 75% identity using BLAST (blastn) and ARG-ANNOT are indicated in green. Antimicrobials to which each isolate is resistant are indicated in red, and intermediate resistance to an antimicrobial is indicated in orange. Plasmid replicons detected in each genome sequence using PlasmidFinder are indicated in purple. Branch lengths are reported in substitutions per site, while posterior probabilities are reported at tree nodes.

MANOVA found significant associations between AMR genes and either source or state after correcting for multiple testing ($P > 0.05$) (Table 2.2), Fisher's exact test indicated that the IncI1 replicon was more commonly detected in New York State isolates than in Washington State isolates (Table 2.5) ($P < 0.05$, after Holm-Bonferroni correction).

Table 2.5: Odds ratios for association of AMR gene groups, AMR phenotype, and plasmid replicons with source or location (only associations with P values of < 0.05 are shown).[a]

Characteristic	*Serotype*	*Source/location favored by OR*	*OR*	*Uncorrected P value*
Source				
Gene				
aac(3)-IIa	Typhimurium	Human	Infinity (only in humans)	0.009
floR	Typhimurium	Human	5.42	0.021
aph(3″)-Ia	Newport	Bovine	0.0831	**0.019**
Antimicrobial				
CHL	Typhimurium	Human	5.42	0.021
NAL	Typhimurium	Human	Infinity (only in humans)	0.022
SXT	Typhimurium	Human	Infinity (only in humans)	**0.004**
TET	Typhimurium	Human	Infinity (all human isolates)	**0.005**
STR	Dublin	Human	9.28	**0.030**
Plasmid				
IncA/C2	Typhimurium	Human	8.18	0.048
ColpVC	Newport	Bovine	0 (found in all bovine isolates)	**0.024**
Geographic location				
Gene				
$bla_{TEM\text{-}1D}$	Typhimurium	WA	4.60	0.045
aph(3″)-Ia	Newport	NY	0.172	**0.049**
aadB	Dublin	WA	Infinity (found only in WA)	**0.005**
cmlA	Dublin	WA	Infinity (found only in WA)	**0.005**
Antimicrobial				
NAL	Typhimurium	WA	Infinity (found only in WA)	0.020
STR	Typhimurium	WA	8.51	0.042
SX	Typhimurium	WA	10.8	0.019
SXT	Typhimurium	WA	9.36	0.042
STR	Dublin	NY	0.052	**0.008**
Plasmid				
IncI1	Typhimurium	NY	0.0602	**0.003**
IncP	Typhimurium	WA	Infinity (found only in WA)	0.046
IncFII(S)	Dublin	NY	0 (present in all NY isolates)	**0.001**

[a] An odds ratio (OR) of infinity or 0 includes a short statement (in parentheses) that indicates which source or location was the driver for that OR (e.g., only in humans indicates that the given gene/phenotype/plasmid replicon was found in only human isolates and in none of the bovine isolates). WA, Washington State; NY, New York State. Values in boldface were significant ($P < 0.05$) after a Holm-Bonferroni correction was applied to the respective analysis.

At the phenotypic level, the number of antimicrobials to which *S.* Typhimurium isolates were resistant ranged from 1 to 11 (Figure 2.3). The most common phenotypic resistance profiles for *S.* Typhimurium were AMC-AMP-CHL-SX-STR-TET and AMC-AMP-FOX-TIO-CRO, which were found in 27% and 11% of the isolates, respectively. When ANOSIM and PERMANOVA

were used as metrics to assess clustering, no significant differences between bovine and human clusters or between New York and Washington State clusters formed by phenotypic resistance/susceptibility profiles were detected ($P > 0.05$ after a Holm-Bonferroni correction [Table 2.2]). However, when Fisher's exact test was used to test for differences at the individual antimicrobial level, resistance to SXT was seen only in human-associated *S.* Typhimurium isolates ($P < 0.05$ after a Holm-Bonferroni correction [Table 2.5]). In addition, all human-associated *S.* Typhimurium isolates were resistant to TET, while only 65% of bovine isolates were resistant to TET ($P < 0.05$ after a Holm-Bonferroni correction [Table 2.5]).

In addition to possessing the most diverse genotypic and phenotypic AMR profiles, *S.* Typhimurium was the only serotype in which resistance to NAL (a quinolone) and CIP (a fluoroquinolone) was observed. All isolates that were resistant to NAL and CIP originated from human clinical samples in Washington State (Figure 2.3). *qnr* genes, which are plasmid-mediated quinolone resistance (PMQR) genes, were detected in the sequences of the two *S.* Typhimurium isolates that showed intermediate resistance to NAL (Table 2.6). For each of the four NAL-resistant isolates, point mutations were identified in the quinolone resistance-determining region (QRDR) of *gyrA* (Table 2.6). These nucleotide changes resulted in non-synonymous amino acid changes (Asp87Asn, Asp87Tyr, and Ser83Tyr) that have been previously observed in quinolone-resistant *Salmonella* isolates (Cloeckaert and Chaslus-Dancla 2001). In addition, three of the four NAL-resistant isolates possessed *oqxA* and *oqxB* (Table 2.6). These genes encode the OqxAB multidrug efflux pump, which confers resistance to multiple agents, including low-level resistance to quinolones (Andres et al. 2013; Hansen et al. 2007).

Table 2.6: *S.* Typhimurium isolates with *qnr* and/or *oqx* genes and/or point mutations in *gyrA* and/or *gyrB* and/or *parC*.[a]

Isolate	S/I/R status			Point mutation[b] detected in:		
	NAL	*CIP*	*qnr and/or oqx gene(s) detected*	*gyrA*	*gyrB*	*parC*
BOV_TYPH_NY_12_R8_9801	S	S	None	1641: T → G	WT	WT
BOV_TYPH_NY_12_R8_9815	S	S	None	1641: T → G	WT	WT
BOV_TYPH_NY_12_R8_9832	S	S	None	1641: T → G	WT	WT
HUM_TYPH_NY_11_R8_8073	S	S	None	WT	2202: G → A	WT
HUM_TYPH_NY_12_R9_0042	S	S	None	WT	2202: G → A	WT
HUM_TYPH_WA_08_R9_3269	I	S	*qnrS*	WT	WT	1713: C → T
HUM_TYPH_WA_08_R9_3270	R	I	*oqxA, oqxB*	Asp87Tyr 259: G → T	WT	1713: C → T
HUM_TYPH_WA_09_R9_3271	S	S	None	WT	759: A → G	WT
HUM_TYPH_WA_10_R9_3273	R	S	*oqxA, oqxB*	Ser83Tyr 248: C → A	WT	1713: C → T
HUM_TYPH_WA_10_R9_3274	I	S	*qnrB*	WT	WT	WT
HUM_TYPH_WA_11_R9_3275	R	S	*oqxA, oqxB*	Asp87Asn 259: G → A	WT	1713: C → T
HUM_TYPH_WA_11_R9_3276	R	S	None	Asp87Asn 259: G → A	WT	1713: C → T
HUM_TYPH_WA_12_R9_3277	S	S	None	WT	WT	1713: C → T

[a] No point mutations were detected in *parE*.
[b] For *gyrA*, *gyrB*, and *parC*, synonymous point mutations resulting in no amino acid change are shown as position: nt → nt (e.g., 259: G → A); amino acid substitutions are formatted as "reference amino acid:position:alternate amino acid"; WT, gene with no mutations.

2.4.4 *S.* Newport phylogeny, AMR genes, AMR phenotypes, and plasmid replicons

Among the 19 *S.* Newport isolates from New York State, 11 clustered into a single, well-supported clade (posterior probability, 1) (Figure 2.4). The inclusion of an additional isolate from New York State yielded a 12-isolate clade with a posterior probability of 0.9574.

The AMR gene profiles of the 32 *S.* Newport isolates showed a high degree of similarity, with only 5 different genotypic profiles (Figure 2.4). The two most common genotypic profiles, i.e., *aac(6)-Iaa floR CMY sul2 tet(A) aph(3")-Ia strB strA tet(R) PBP* gene and *aac(6)-Iaa floR CMY sul2 tet(A) strB strA tet(R) PBP* gene, were detected in 66% and 19% of *S.* Newport genomes, respectively. At the individual gene level, genes belonging to the *aac(6)-Iaa*, *CMY*, *strA*, *strB*, *sul2*, *tet(A)*, *tet(R)*, and *PBP* gene groups were detected in the sequences of all 32 isolates (Table 2.1). All *S.* Newport isolates had identical copies of each of these genes except for *CMY*, as a truncated version of the gene was detected in isolate

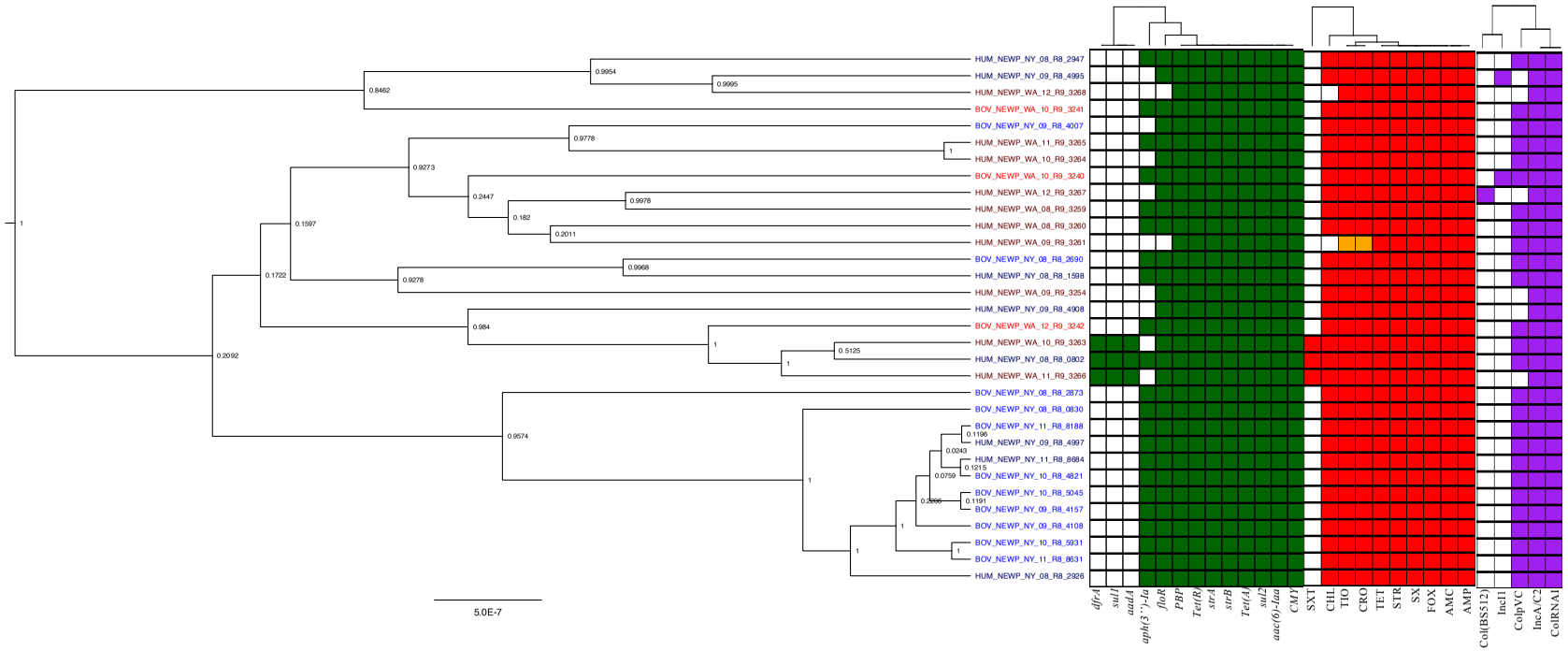


Figure 2.4: Phylogenetic tree of *S.* Newport isolates constructed using BEAST. Gene groups for AMR genes detected in each genome sequence at more than 50% coverage and 75% identity using BLAST (blastn) and ARG-ANNOT are indicated in green. Antimicrobials to which each isolate is resistant are indicated in red, and intermediate resistance to an antimicrobial is indicated in orange. Plasmid replicons detected in each genome sequence using PlasmidFinder are indicated in purple. Branch lengths are reported in substitutions per site, while posterior probabilities are reported at tree nodes.

BOV_NEWP_WA_10_R9_3240. In addition, the IncA/C2 and ColRNAI replicons were detected in all *S.* Newport genomes (Table 2.1). Neither ANOSIM nor PERMANOVA detected significant associations between AMR genes or plasmid replicon presence/absence and source after correcting for multiple testing ($P > 0.05$ after a Holm-Bonferroni correction [Table 2.2]). However, the AMR gene sequences of Washington State and New York State isolates were found to differ when ANOSIM was used as a metric ($P < 0.05$ after a Holm-Bonferroni correction [Table 2.2]). When Fisher's exact test was used to assess source and geographic associations at the individual gene level, genes belonging to the *aph(3")-Ia* group were more commonly present in (i) *S.* Newport bovine isolates and (ii) isolates from New York State ($P < 0.05$ after a Holm-Bonferroni correction [Table 2.5]). Additionally, the ColpVC plasmid replicon was detected in all bovine *S.* Newport isolates and only 67% of the human isolates ($P < 0.05$ after a Holm-Bonferroni correction [Table 2.5]).

S. Newport isolates appeared even more similar at the phenotypic AMR level than at the genetic level. No significant source or geographic differences in MDR phenotype were observed when ANOSIM and PERMANOVA were used to assess clustering ($P > 0.05$ after a Holm-Bonferroni correction) (Table 2.2). All 32 *S.* Newport isolates were resistant to AMC, AMP, FOX, TIO, CRO, SX, STR, and TET, and only 3 different phenotypic profiles were detected (Figure 2.4). The most common of these, AMC-AMP-FOX-TIO-CRO-CHL-SX-STR-TET, was carried by 27 of the 32 (84%) *S.* Newport isolates. Three isolates showed additional resistance to SXT; hence, the two most common profiles accounted for 30 of the 32 (94%) isolates. The three SXT-resistant isolates possessed *aadA*, *dfrA*, and *sul1*, which were not detected in any other *S.* Newport genomes (Figure 2.4).

2.4.5 *S.* Dublin phylogeny, AMR genes, AMR phenotypes, and plasmid replicons

S. Dublin isolates clustered into two separate clades with a posterior probability of 1, one of which consisted of 10 isolates exclusively from Washington State (referred to here as the Washington State clade) (Figure 2.5). The other clade included all eight *S.* Dublin isolates from New York State and three isolates from Washington State (referred to here as the mixed clade) (Figure 2.5). Both genotypic and phenotypic differences were observed between the two major clades. AMR genes *aadB* and *cmlA*, which were detected in all but 1 Washington State state clade isolate, were not detected in any of the mixed clade isolates ($P < 0.05$ after a Holm-Bonferroni correction) (Figure 2.5). Not surprisingly, the frequencies at which these genes were detected in New York and Washington States were significantly different when Fisher's exact test was used ($P < 0.05$ after a Holm-Bonferroni correction) (Table 2.5). ANOSIM and PERMANOVA did not identify significant differences between *S.* Dublin geographic clusters formed by AMR gene sequences (Table 2.2). However, when ANOSIM and PERMANOVA were conducted using plasmid replicon presence/absence data, significant differences between New York and Washington State isolate clusters were observed for *S.* Dublin ($P < 0.05$ after a Holm-Bonferroni correction) (Table 2.2). In addition, when Fisher's exact test was used to test for possible geographic associations of individual plasmid replicons, the IncFII(S) replicon was detected only in mixed clade isolates, making it more commonly associated with isolates from New York State ($P < 0.05$ after a Holm-Bonferroni correction) (Figure 2.5).

Significant differences between New York and Washington State isolate clus-

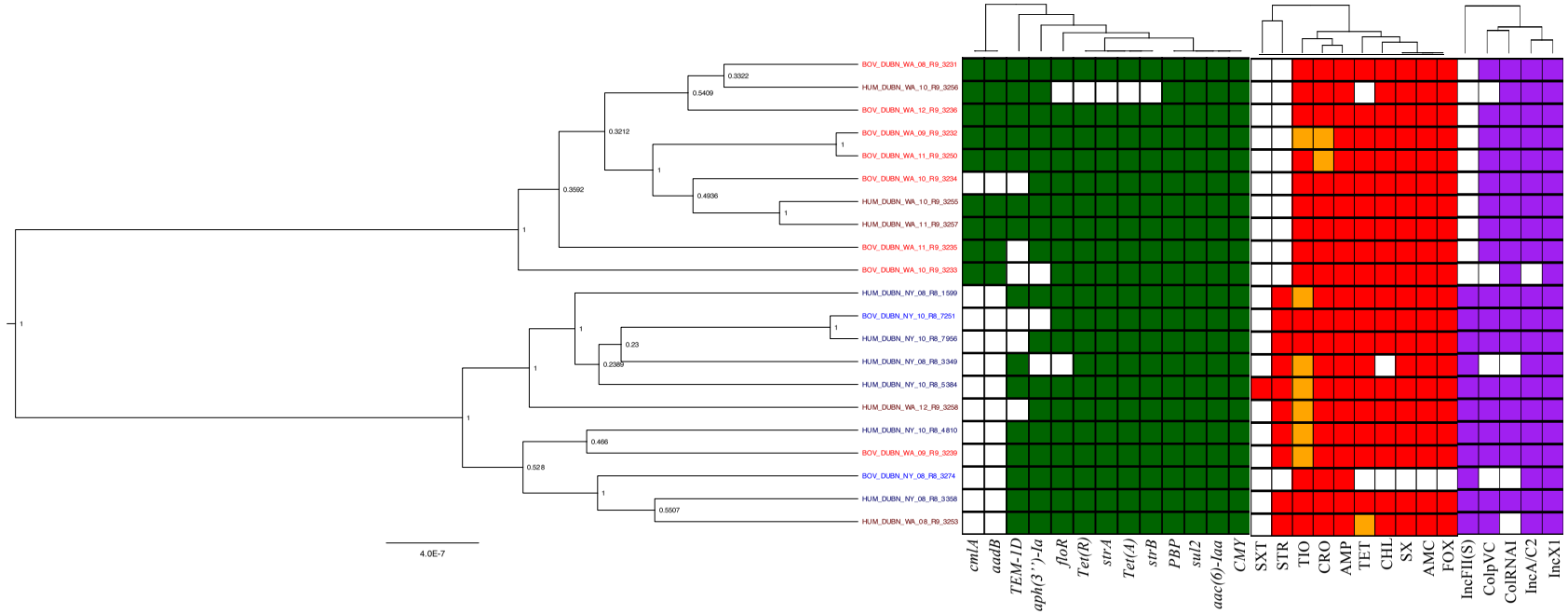


Figure 2.5: Phylogenetic tree of *S.* Dublin isolates constructed using BEAST. Gene groups for AMR genes detected in each genome sequence at more than 50% coverage and 75% identity using BLAST (blastn) and ARG-ANNOT are indicated in green. Antimicrobials to which each isolate is resistant are indicated in red, and intermediate resistance to an antimicrobial is indicated in orange. Plasmid replicons detected in each genome sequence using PlasmidFinder are indicated in purple. Branch lengths are reported in substitutions per site, while posterior probabilities are reported at tree nodes.

ters were observed for *S.* Dublin when ANOSIM and PERMANOVA were conducted using phenotypic resistance/susceptibility data ($P < 0.05$ after a Holm-Bonferroni correction) (Table 2.2). Despite the detection of both *strA* and *strB* in 20 of the 21 genomes (Table 2.1), STR resistance was observed only in isolates in the mixed clade ($P < 0.05$ after a Holm-Bonferroni correction) (Figure 2.5). While the *strB* sequence was the same for the 20 isolates, the *strA* sequence showed a strong geographical association: all isolates in the Washington State clade possessed a truncated form of the gene, with the first 91 bp of the gene missing. Aside from this 91-bp deletion, the *strA* sequences were identical in all isolates. Overall, 11 isolates carried *strB* and a full-length *strA*; 10 of these isolates showed phenotypic STR resistance. However, 9 isolates carried *strB* and a truncated *strA*; all of these isolates were sensitive to STR. These data suggest that the presence of the truncated *strA* variant found here does not confer STR resistance and also suggest that the presence of only the *strB* variant found here, in the absence of a full-length *strA*, does not confer STR resistance.

The *S.* Dublin isolates were distributed into 8 different AMR genotypic profiles, with 33% of isolates genes belonging to the *aac(6)-Iaa floR CMY sul2 tet(A) aph(3")-Ia blaTEM-1D strB strA tet(R) PBP* gene genotypic profile. The most common resistance genes in *S.* Dublin belonged to the *aac(6)-Iaa*, *CMY*, and *sul2* groups, all of which were detected in all 21 *S.* Dublin isolates (Table 2.1). The sequences of these genes were identical for all *S.* Dublin isolates, regardless of source or geographic location. The *PBP* gene was also detected in all 21 genomes (Table 2.1). *PBP* gene sequences for 20 isolates were identical; only the sequence of isolate BOV_DUBN_WA_09_R9_3239 differed by a single nucleotide from the 20 other sequences. In addition, the replicon for IncX1, which had been detected in only 1 *S.* Typhimurium isolate and no *S.* Newport isolates in

this study, was detected in all 21 *S.* Dublin genomes (Figure 2.5). At the phenotypic level, 6 different phenotypic profiles were observed. The two most common, AMC-AMP-FOX-TIO-CRO-CHL-SX-TET and AMC-AMP-FOX-TIO-CRO-CHL-SX-STR-TET, were observed in 43% and 38% of *S.* Dublin isolates, respectively.

2.5 Discussion

Antimicrobial resistance in zoonotic and foodborne pathogens is considered to be one of the most serious threats to public health today (CDC 2013; WHO 2014). The emergence and dispersal of AMR *Salmonella* are particularly problematic, due to (i) the fact that non-typhoidal *Salmonella* represents one of the most common causes of foodborne disease cases and associated deaths worldwide (WHO 2015) and (ii) reports on the emergence and dispersal of different multidrug-resistant *Salmonella* strains (e.g., *Salmonella* Typhimurium DT104) (Helms et al. 2005; Leekitcharoenphon et al. 2016; Ribot et al. 2002). Studies of the relationships between AMR determinants and MDR strains found in humans and animals are often confounded by the selection of the isolates included in a given study, in which human and animal isolates may be of different serotypes, geographical locations, or temporal intervals. To further our understanding of AMR diversity and dispersal in *Salmonella*, we thus assembled and characterized a set of *Salmonella* isolates that (i) represented 3 serotypes associated with both human and bovine populations, (ii) were isolated over the same time frame (2008 to 2012), (iii) were matched by source (human or animal) so that approximately equal numbers of human and bovine isolates were selected from each serotype, and (iv) were matched by geographical location so that sim-

ilar numbers of human and bovine isolates of the three different serotypes were obtained from each of the states of Washington and New York. Our data obtained from these isolates suggest that (i) WGS can be used to reliably predict phenotypic resistance across *Salmonella* isolates from both human and bovine sources, (ii) geographical differences can contribute to distinct, location-specific AMR patterns, and (iii) despite an overlap of AMR geno- and phenotypes, human and bovine isolates differ significantly based on a number of AMR-related geno- and phenotypic characteristics.

2.5.1 WGS can be used to predict phenotypic resistance in bovine and human-associated *Salmonella* Typhimurium, Newport, and Dublin with high sensitivity and specificity

Our study reported here demonstrates that *in silico* AMR gene predictions are highly correlated with phenotypic resistance in *Salmonella enterica* Typhimurium, Newport, and Dublin, as AMR genotype correlated with AMR phenotype with an overall sensitivity and specificity of 97.2 and 85.2%, respectively. The ability to predict AMR phenotype from WGS data with high sensitivity and specificity has previously been observed in *Salmonella enterica* isolated from humans and retail meats (McDermott et al. 2016) and *S.* Typhimurium from swine (Zankari et al. 2012), as well as in other organisms, including *Staphylococcus aureus* (Gordon et al. 2014; Bradley et al. 2015), *Campylobacter* spp. (Zhao et al. 2016), and *Mycobacterium tuberculosis* (Bradley et al. 2015). The results of our study further attest to the robustness of WGS in predicting resistance phenotypes in *Salmonella enterica* serotypes Typhimurium, Newport, and Dublin from

both bovine and human sources. Verification of the ability of WGS to predict phenotypic AMR in bovine isolates is important, as AMR in isolates from different hosts can be facilitated by different mechanisms, as also shown here. Our data further support that as WGS becomes faster, cheaper, and more accessible, it may represent a valuable tool that could replace classical phenotypic AMR testing across human medical, public health, and veterinary fields.

In this study, the lowest sensitivity of predicting AMR phenotype from genotypic data occurred for NAL. This was not surprising, since the AMR phenotype prediction approach used here was based on the presence of genes that confer resistance to a given antibiotic. While AMR gene-based approaches generally work well, quinolone and fluoroquinolone resistance in particular can result from point mutations in housekeeping genes (e.g., *gyrA*) rather than from the presence of resistance genes, even though the presence of some resistance genes (e.g., PMQR genes) may also confer low-level resistance to quinolones and fluoroquinolones (Cloeckaert and Chaslus-Dancla 2001; Hooper and Jacoby 2015). In our study, the two isolates that showed intermediate resistance to NAL possessed PMQR genes, but no mutations in housekeeping genes known to confer resistance to quinolones. This is consistent with previous findings, in which isolates possessing PMQR genes have been shown to have reduced susceptibility to quinolones but were not clinically resistant (Hooper and Jacoby 2015). Of the four NAL-resistant isolates, three concurrently possessed PMQR genes and non-synonymous mutations in the quinolone resistance-determining region (QRDR) of *gyrA*. One isolate that was NAL resistant due to the presence of only a non-synonymous mutation in *gyrA* was falsely predicted to be NAL sensitive, due to an absence of quinolone resistance genes in its genome. This showcases that relying solely on gene presence/absence to predict AMR can re-

sult in reduced sensitivity. However, this drawback can be easily alleviated by incorporating SNP-based prediction of AMR (as now has been implemented in the ARG-ANNOT and CARD bioinformatic tools) (Gupta et al. 2014; Jia et al. 2017).

In this study, the lowest specificity of WGS-based AMR prediction was observed for STR, which accounted for more than one-half of all phenotype-susceptible/genotype-resistant (P^-:G^+) discrepancies. Here, more than 50% of these discrepancies were attributed to *S.* Dublin isolates from the Washington State clade, which carry a truncated *strA* that appeared to not confer STR resistance, while still being identified computationally as an STR resistance determinant. Similar discrepancies have been observed in a previous study (Davis, Besser, Orfe, et al. 2011) of *Escherichia coli* isolates from dairy calves; in this study, point mutations in *strA* were hypothesized to affect its ability to confer STR resistance. Additionally, a previous study that assessed phenotypic and genotypic resistance in non-typhoidal *Salmonella* isolated from retail meat and human clinical samples also found STR (P^-:G^+) discrepancies to be the most common (McDermott et al. 2016). The authors of this previous study suggest that STR (P^-:G^+) discrepancies could be due to inaccurate clinical breakpoints for STR susceptibility in *Salmonella*, due in part to the fact that STR is not used to treat enteric infections (McDermott et al. 2016). Overall, these findings suggest that refinement of WGS-based AMR prediction methods could benefit from the incorporation of tools that also classify specific allelic variants of resistance genes for their ability (or inability) to confer resistance. In the future, WGS-based AMR prediction tools that incorporate feedback from clinical use of antibiotics may even further improve the ability of WGS-based tools to predict the clinical outcome of treatment with a given antimicrobial.

2.5.2 Both phenotypic and genomic data show geographic differences in resistance-related characteristics for *Salmonella*, suggesting a need for location-specific AMR control strategies.

Our data show significant differences between New York and Washington State isolates with regard to AMR-relevant genotypic and phenotypic characteristics. Specifically, when ANOSIM and/or PERMANOVA were used as metrics, Washington and New York State isolates differed by (i) AMR gene sequences (in serotype Newport) and (ii) phenotypic resistance/susceptibility and plasmid replicon presence/absence (in serotype Dublin) (Table 2.2). In addition, a number of genes, antimicrobials, and plasmid replicons showed strong geographical associations, even after corrections for multiple testing (Table 2.5). For example, the presence of *aadB* and *cmlA* was associated with *S.* Dublin isolates from Washington State, while STR resistance was associated with *S.* Dublin from New York State. In *S.* Typhimurium, the IncI1 plasmid replicon, which has been previously associated with extended-spectrum cephalosporin resistance in *S.* Typhimurium (Folster et al. 2014; Jean-Yves Madec et al. 2011), was more commonly detected in isolates from New York State. In *S.* Dublin, the IncFII(S) plasmid replicon was also more commonly detected in isolates from New York State; the IncFII(S) replicon, along with IncFIB(S), are characteristic of the *Salmonella* virulence plasmids (Carattoli et al. 2014) found in serotypes such as *S.* Typhimurium and *S.* Dublin, and it has been proposed that some virulence plasmids previously associated with *S.* Dublin have evolved from IncFII-like plasmids (Chu et al. 2008). The geographic differences observed for

MDR-relevant genotypic and phenotypic characteristics suggest that different ecological factors and selective pressures may contribute to the development of AMR in different geographical locations (New York State and Washington State in our study here), suggesting a need for geographically specific interventions to effectively combat the spread of AMR. Our findings are consistent with previous studies that have shown that contemporary *Salmonella* antibiotic resistance patterns differ, even within a given country. For example, Davis et al. (Davis, Besser, Eckmann, et al. 2007) showed that a specific MDR *Salmonella* Typhimurium strain emerged prior to 2000 in bovine populations in the Pacific Northwest (which includes Washington State) but was not found among contemporary isolates from the Northeast. Similarly, a large-scale WGS study of *Salmonella* Typhi isolates from across the world identified a specific MDR clone that emerged in Asia and Africa with subsequent inter- and intracontinental transmission events (Wong et al. 2015). Importantly, our findings are also consistent with a WGS-based study (Strachan et al. 2015) of *Escherichia coli* O157 isolates from different sources (e.g., animals, humans, and the environment/food) and different countries and continents. This study reported significant genetic differences among isolates from different geographical regions and hypothesized that a combination of local emergence events and international transmission leads to a "patchwork" of geographically confined and widely distributed clades. This is similar to what we have observed, as we have identified certain geographic location-specific clones (e.g., a Washington State-specific Dublin clade that carries a truncated *strA* allele), as well as broadly distributed clonal groups with similar AMR profiles.

2.5.3 *S. enterica* isolates from humans contain a more diverse range of AMR genes and plasmid replicons than those isolated from bovine populations

The development and spread of AMR have often been attributed to the misuse of antimicrobials in agricultural settings. However, the AMR profiles of *Salmonella* isolated from human infections cannot be fully explained by AMR in bovine isolates in this study alone. Here, resistance to CIP, NAL, and SXT were observed only in isolates from humans with salmonellosis. At the genotypic level, over one-half of the total of 42 AMR genes detected in this study were detected only in human isolates. Similar results were observed for plasmid replicons, as nearly one-half of the plasmid replicons detected were found only in human isolates. These results, along with the phylogenetic relationship of the isolates, suggest that some AMR genes are associated primarily with a particular host, with little overlap between species. Mather et al. (A. E. Mather et al. 2013) observed similar results for human- and animal-associated *S.* Typhimurium DT104: *Salmonella* isolates from humans and animals, as well as the AMR genes associated with them, were found to remain largely within their respective host populations, with little transmission from animals to humans and vice versa (A. E. Mather et al. 2013).

While many AMR genes and phenotypes were confined to the human isolates in this study, overlaps between the resistomes of bovine and human-associated *Salmonella* isolates were observed on numerous occasions, with the high degree of AMR sequence identity observed for *S.* Newport isolates serving as the most prominent example. This also is consistent with previous studies

(Spoor et al. 2013; Ward et al. 2014; J.-Y. Madec et al. 2017) that similarly described that certain clonal groups of AMR pathogens can be found in both humans and animals. However, further studies using WGS data from temporally sampled *Salmonella enterica* are needed to assess the spread of AMR *Salmonella* and the resistance genes associated with it in New York State and Washington State.

2.6 Acknowledgments

This material is based on work supported by the National Science Foundation Graduate Research Fellowship Program under grant no. DGE-1144153. Research reported in this publication was supported by the Agriculture and Food Research Initiative Competitive Grant no. 2010-51110-21131 from the USDA National Institute of Food and Agriculture. The content is solely the responsibility of the authors and does not necessarily represent the official views of the USDA.

2.7 References

Anderson, Marti J. (2001). "A new method for non-parametric multivariate analysis of variance". In: *Austral Ecology* 26.1, pp. 32–46. DOI: doi:10.1111/j.1442-9993.2001.01070.pp.x.

Anderson, Marti J. and Daniel C. I. Walsh (2013). "PERMANOVA, ANOSIM, and the Mantel test in the face of heterogeneous dispersions: What null hypothesis are you testing?" In: *Ecological Monographs* 83.4, pp. 557–574. DOI: 10.1890/12-2010.1. eprint: https://esajournals.onlinelibrary.wiley.com/doi/pdf/10.1890/12-2010.1.

Andres, Patricia et al. (2013). "Differential distribution of plasmid-mediated quinolone resistance genes in clinical enterobacteria with unusual phenotypes of quinolone susceptibility from Argentina". In: *Antimicrob Agents Chemother* 57.6, pp. 2467–2475. DOI: 10.1128/AAC.01615-12.

Andrews, S. (2014). "FastQC A Quality Control tool for High Throughput Sequence Data". In: *http://www.bioinformatics.babraham.ac.uk/projects/fastqc/*. DOI: citeulike-article-id:11583827.

Baele, Guy, Philippe Lemey, et al. (2012). "Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty". In: *Mol Biol Evol* 29.9, pp. 2157–2167. DOI: 10.1093/molbev/mss084.

Baele, Guy, Wai Lok Sibon Li, Alexei J. Drummond, Marc A. Suchard, and Philippe Lemey (2013). "Accurate model selection of relaxed molecular clocks in bayesian phylogenetics". In: *Mol Biol Evol* 30.2, pp. 239–243. DOI: 10.1093/molbev/mss243.

Bankevich, A. et al. (2012). "SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing". In: *J Comput Biol* 19.5, pp. 455–77. DOI: 10.1089/cmb.2012.0021.

Bolger, A. M., M. Lohse, and B. Usadel (2014). "Trimmomatic: a flexible trimmer for Illumina sequence data". In: *Bioinformatics* 30.15, pp. 2114–20. DOI: 10.1093/bioinformatics/btu170.

Bradley, Phelim et al. (2015). "Rapid antibiotic-resistance predictions from genome sequence data for *Staphylococcus aureus* and *Mycobacterium tuberculosis*". In: *Nat Commun* 6, pp. 10063–10063. DOI: 10.1038/ncomms10063.

Bushnell, B. (2015). "BBMap v. 35.49, https://sourceforge.net/projects/bbmap/". In:

Camacho, C. et al. (2009). "BLAST+: architecture and applications". In: *BMC Bioinformatics* 10, p. 421. DOI: 10.1186/1471-2105-10-421.

Carattoli, A. et al. (2014). "*In silico* detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing". In: *Antimicrob Agents Chemother* 58.7, pp. 3895–903. DOI: 10.1128/AAC.02412-14.

CDC (2013). *Antibiotic resistance threats in the United States, 2013*. CDC, Atlanta, GA.

Chase, Jonathan M., Nathan J. B. Kraft, Kevin G. Smith, Mark Vellend, and Brian D Inouye (2011). "Using null models to disentangle variation in community dissimilarity from variation in alpha-diversity". In: *Ecosphere* 2.2, art24. DOI: 10.1890/ES10-00117.1. eprint: https://esajournals.onlinelibrary.wiley.com/doi/pdf/10.1890/ES10-00117.1.

Chu, Chishih et al. (2008). "Evolution of genes on the *Salmonella* Virulence plasmid phylogeny revealed from sequencing of the virulence plasmids of *S. enterica* serotype Dublin and comparative analysis". In: *Genomics* 92.5, pp. 339–343.

Clarke, K. R. (1993). "Non-parametric multivariate analyses of changes in community structure". In: *Australian Journal of Ecology* 18.1, pp. 117–143. DOI: 10.1111/j.1442-9993.1993.tb00438.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1442-9993.1993.tb00438.x.

Cloeckaert, Axel and Elisabeth Chaslus-Dancla (2001). "Mechanisms of quinolone resistance in *Salmonella*". In: *Vet. Res.* 32.3-4, pp. 291–300. DOI: 10.1051/vetres:2001105.

CLSI (2012). *Performance standards for antimicrobial susceptibility testing, twenty-second informational supplement. M100-D22, 22nd ed*. Clinical and Laboratory Standards Institute, Wayne, PA.

— (2013). *Performance standards for antimicrobial disk and dilution susceptibility tests for bacteria isolated from animals approved standard, fourth edition, VET01-A4, 3rd ed*. Clinical and Laboratory Standards Institute, Wayne, PA.

Cody, Sara H. et al. (1999). "Two Outbreaks of Multidrug-Resistant *Salmonella* Serotype Typhimurium DT104 Infections Linked to Raw-Milk Cheese in

Northern California". In: *JAMA* 281.19, pp. 1805–1810. DOI: 10.1001/jama.281.19.1805. eprint: https://jamanetwork.com/journals/jama/articlepdf/189982/joc81201.pdf.

Croucher, N. J. et al. (2015). "Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins". In: *Nucleic Acids Res* 43.3, e15. DOI: 10.1093/nar/gku1196.

Davis, Margaret A., Thomas E. Besser, Kaye Eckmann, et al. (2007). "Multidrug-resistant *Salmonella* typhimurium, Pacific Northwest, United States". In: *Emerg Infect Dis* 13.10, pp. 1583–1586. DOI: 10.3201/eid1310.070536.

Davis, Margaret A., Thomas E. Besser, Lisa H. Orfe, et al. (2011). "Genotypic-Phenotypic Discrepancies between Antibiotic Resistance Characteristics of *Escherichia coli* Isolates from Calves in Management Settings with High and Low Antibiotic Use". In: *Applied and Environmental Microbiology* 77.10, pp. 3293–3299. DOI: 10.1128/AEM.02588-10. eprint: https://aem.asm.org/content/77/10/3293.full.pdf.

Drummond, A. J., S. Y. Ho, M. J. Phillips, and A. Rambaut (2006). "Relaxed phylogenetics and dating with confidence". In: *PLoS Biol* 4.5, e88. DOI: 10.1371/journal.pbio.0040088.

Drummond, A. J., A. Rambaut, B. Shapiro, and O. G. Pybus (2005). "Bayesian coalescent inference of past population dynamics from molecular sequences". In: *Mol Biol Evol* 22.5, pp. 1185–92. DOI: 10.1093/molbev/msi103.

Drummond, Alexei J., Marc A. Suchard, Dong Xie, and Andrew Rambaut (2012). "Bayesian phylogenetics with BEAUti and the BEAST 1.7". In: *Mol Biol Evol* 29.8, pp. 1969–1973. DOI: 10.1093/molbev/mss075.

Fey, Paul D. et al. (2000). "Ceftriaxone-Resistant *Salmonella* Infection Acquired by a Child from Cattle". In: *New England Journal of Medicine* 342.17. PMID: 10781620, pp. 1242–1249. DOI: 10.1056/NEJM200004273421703. eprint: https://doi.org/10.1056/NEJM200004273421703.

Folster, Jason P. et al. (2014). "Characterization of *blaCMY* plasmids and their possible role in source attribution of *Salmonella enterica* serotype Ty-

phimurium infections". In: *Foodborne Pathog Dis* 11.4, pp. 301–306. DOI: 10.1089/fpd.2013.1670.

Fricke, W. Florian et al. (2009). "Comparative genomics of the IncA/C multidrug resistance plasmid family". In: *J Bacteriol* 191.15, pp. 4750–4757. DOI: 10.1128/JB.00189-09.

Gardner, S. N. and B. G. Hall (2013). "When whole-genome alignments just won't work: kSNP v2 software for alignment-free SNP discovery and phylogenetics of hundreds of microbial genomes". In: *PLoS One* 8.12, e81760. DOI: 10.1371/journal.pone.0081760.

Gordon, N. C. et al. (2014). "Prediction of *Staphylococcus aureus* Antimicrobial Resistance by Whole-Genome Sequencing". In: *Journal of Clinical Microbiology* 52.4. Ed. by K. C. Carroll, pp. 1182–1191. DOI: 10.1128/JCM.03117-13. eprint: https://jcm.asm.org/content/52/4/1182.full.pdf.

Gupta, S. K. et al. (2014). "ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes". In: *Antimicrob Agents Chemother* 58.1, pp. 212–20. DOI: 10.1128/AAC.01310-13.

Hald, Tine et al. (2016). "World Health Organization Estimates of the Relative Contributions of Food to the Burden of Disease Due to Selected Foodborne Hazards: A Structured Expert Elicitation". In: *PLOS ONE* 11.1, pp. 1–35. DOI: 10.1371/journal.pone.0145839.

Hansen, Lars Hestbjerg, Lars Bogo Jensen, Heidi Iskou Sorensen, and Soren Johannes Sorensen (2007). "Substrate specificity of the OqxAB multidrug resistance pump in *Escherichia coli* and selected enteric bacteria". In: *Journal of Antimicrobial Chemotherapy* 60.1, pp. 145–147. DOI: 10.1093/jac/dkm167. eprint: http://oup.prod.sis.lan/jac/article-pdf/60/1/145/2178195/dkm167.pdf.

Helms, M., S. Ethelberg, K. Molbak, and D. T. Study Group (2005). "International *Salmonella* Typhimurium DT104 infections, 1992-2001". In: *Emerg Infect Dis* 11.6, pp. 859–67. DOI: 10.3201/eid1106.041017.

Hendriksen, Susan W. M., Karin Orsel, Jaap A. Wagenaar, Angelika Miko, and Engeline van Duijkeren (2004). "Animal-to-human transmission of *Salmonella* Typhimurium DT104A variant". In: *Emerg Infect Dis* 10.12, pp. 2225–2227. DOI: 10.3201/eid1012.040286.

Hoelzer, Karin, Andrea Isabel Moreno Switt, and Martin Wiedmann (2011). "Animal contact as a source of human non-typhoidal salmonellosis". In: *Vet Res* 42.1, pp. 34–34. DOI: 10.1186/1297-9716-42-34.

Holm, Sture (1979). "A Simple Sequentially Rejective Multiple Test Procedure". In: *Scandinavian Journal of Statistics* 6.2, pp. 65–70.

Holmes, A. et al. (2015). "Utility of Whole-Genome Sequencing of *Escherichia coli* O157 for Outbreak Detection and Epidemiological Surveillance". In: *J Clin Microbiol* 53.11, pp. 3565–73. DOI: 10.1128/JCM.01066-15.

Hooper, David C. and George A. Jacoby (2015). "Mechanisms of drug resistance: quinolone resistance". In: *Ann N Y Acad Sci* 1354.1, pp. 12–31. DOI: 10.1111/nyas.12830.

Inouye, M. et al. (2014). "SRST2: Rapid genomic surveillance for public health and hospital microbiology labs". In: *Genome Med* 6.11, p. 90. DOI: 10.1186/s13073-014-0090-6.

Iqbal, Zamin, Mario Caccamo, Isaac Turner, Paul Flicek, and Gil McVean (2012). "*De novo* assembly and genotyping of variants using colored de Bruijn graphs". In: *Nature Genetics* 44, pp. 226–232.

Jia, Kun et al. (2017). "Preliminary Transcriptome Analysis of Mature Biofilm and Planktonic Cells of *Salmonella* Enteritidis Exposure to Acid Stress". In: *Front Microbiol* 8, pp. 1861–1861. DOI: 10.3389/fmicb.2017.01861.

Johnson, James R. et al. (2007). "Antimicrobial drug-resistant *Escherichia coli* from humans and poultry products, Minnesota and Wisconsin, 2002-2004". In: *Emerg Infect Dis* 13.6, pp. 838–846. DOI: 10.3201/eid1306.061576.

Kimura, M. (1980). "A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences". In: *J Mol Evol* 16.2, pp. 111–120.

Kruskal, J. B. (1964a). "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis". In: *Psychometrika* 29.1, pp. 1–27. DOI: 10.1007/BF02289565.

— (1964b). "Nonmetric multidimensional scaling: A numerical method". In: *Psychometrika* 29.2, pp. 115–129. DOI: 10.1007/BF02289694.

Kwong, J. C. et al. (2016). "Prospective Whole-Genome Sequencing Enhances National Surveillance of *Listeria monocytogenes*". In: *J Clin Microbiol* 54.2, pp. 333–42. DOI: 10.1128/JCM.02344-15.

Leekitcharoenphon, P. et al. (2016). "Global Genomic Epidemiology of *Salmonella enterica* Serovar Typhimurium DT104". In: *Appl Environ Microbiol* 82.8, pp. 2516–26. DOI: 10.1128/AEM.03821-15.

Li, H. et al. (2009). "The Sequence Alignment/Map format and SAMtools". In: *Bioinformatics* 25.16, pp. 2078–9. DOI: 10.1093/bioinformatics/btp352.

Lozupone, Catherine and Rob Knight (2005). "UniFrac: a new phylogenetic method for comparing microbial communities". In: *Appl Environ Microbiol* 71.12, pp. 8228–8235. DOI: 10.1128/AEM.71.12.8228-8235.2005.

Madec, Jean-Yves, Benoit Doublet, Cecile Ponsin, Axel Cloeckaert, and Marisa Haenni (2011). "Extended-spectrum beta-lactamase *blaCTX-M-1* gene carried on an IncI1 plasmid in multidrug-resistant *Salmonella enterica* serovar Typhimurium DT104 in cattle in France". In: *Journal of Antimicrobial Chemotherapy* 66.4, pp. 942–944. DOI: 10.1093/jac/dkr014. eprint: http://oup.prod.sis.lan/jac/article-pdf/66/4/942/2160001/dkr014.pdf.

Madec, J.-Y., M. Haenni, P. Nordmann, and L. Poirel (2017). "Extended-spectrum β-lactamase/AmpC- and carbapenemase-producing *Enterobacteri-*

aceae in animals: a threat for humans?" In: *Clinical Microbiology and Infection* 23.11, pp. 826–833. DOI: 10.1016/j.cmi.2017.01.013.

Mather, A. E. et al. (2013). "Distinguishable epidemics of multidrug-resistant *Salmonella* Typhimurium DT104 in different hosts". In: *Science* 341.6153, pp. 1514–7. DOI: 10.1126/science.1240578.

Mather, Alison E. et al. (2012). "An ecological approach to assessing the epidemiology of antimicrobial resistance in animal and human populations". In: *Proc Biol Sci* 279.1733, pp. 1630–1639. DOI: 10.1098/rspb.2011.1975.

McDermott, Patrick F. et al. (2016). "Whole-Genome Sequencing for Detecting Antimicrobial Resistance in Nontyphoidal *Salmonella*". In: *Antimicrobial Agents and Chemotherapy* 60.9, pp. 5515–5520. DOI: 10.1128/AAC.01030-16. eprint: https://aac.asm.org/content/60/9/5515.full.pdf.

Oksanen, Jari et al. (2017). *vegan: Community Ecology Package*. R package version 2.4-2.

PLINK/Seq (2014). "PLINK/Seq v. 0.10. https://atgu.mgh.harvard.edu/plinkseq/". In:

Price, Lance B. et al. (2012). "*Staphylococcus aureus* CC398: Host Adaptation and Emergence of Methicillin Resistance in Livestock". In: *mBio* 3.1. Ed. by Fernando Baquero. DOI: 10.1128/mBio.00305-11. eprint: https://mbio.asm.org/content/3/1/e00305-11.full.pdf.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.

Rambaut, A. (2013). *Analysis of variable sites only in BEAST or MrBayes*. https://groups.google.com/forum/#!topic/beast-users/V5vRghILMfw.

Rambaut, A., T. T. Lam, L. Max Carvalho, and O. G. Pybus (2016). "Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen)". In: *Virus Evol* 2.1, vew007. DOI: 10.1093/ve/vew007.

Ribot, Efrain M., Rachel K. Wierzba, Frederick J. Angulo, and Timothy J. Barrett (2002). "*Salmonella enterica* serotype Typhimurium DT104 isolated from humans, United States, 1985, 1990, and 1995". In: *Emerg Infect Dis* 8.4, pp. 387–391. DOI: 10.3201/eid0804.010202.

Scallan, E. et al. (2011). "Foodborne illness acquired in the United States–major pathogens". In: *Emerg Infect Dis* 17.1, pp. 7–15. DOI: 10.3201/eid1701.P1110110.3201/eid1701.091101p1.

Silbergeld, Ellen K., Jay Graham, and Lance B. Price (2008). "Industrial Food Animal Production, Antimicrobial Resistance, and Human Health". In: *Annual Review of Public Health* 29.1. PMID: 18348709, pp. 151–169. DOI: 10.1146/annurev.publhealth.29.020907.090904. eprint: https://doi.org/10.1146/annurev.publhealth.29.020907.090904.

Spoor, Laura E. et al. (2013). "Livestock Origin for a Human Pandemic Clone of Community-Associated Methicillin-Resistant *Staphylococcus aureus*". In: *mBio* 4.4. Ed. by Fernando Baquero. DOI: 10.1128/mBio.00356-13. eprint: https://mbio.asm.org/content/4/4/e00356-13.full.pdf.

Strachan, Norval J. C. et al. (2015). "Whole Genome Sequencing demonstrates that Geographic Variation of *Escherichia coli* O157 Genotypes Dominates Host Association". In: *Scientific Reports* 5. Article, p. 14145.

Tamura, Koichiro, Glen Stecher, Daniel Peterson, Alan Filipski, and Sudhir Kumar (2013). "MEGA6: Molecular Evolutionary Genetics Analysis version 6.0". In: *Mol Biol Evol* 30.12, pp. 2725–2729. DOI: 10.1093/molbev/mst197.

Tavare, Simon. "Some probabilistic and statistical problems in the analysis of DNA sequences". In: *Lectures on mathematics in the life sciences* 17.2, pp. 57–86.

Taylor, Angela J. et al. (2015). "Characterization of Foodborne Outbreaks of *Salmonella enterica* Serovar Enteritidis with Whole-Genome Sequencing Single Nucleotide Polymorphism-Based Analysis for Surveillance and Outbreak Detection". In: *Journal of Clinical Microbiology* 53.10. Ed. by D. J. Diekema, pp. 3334–3340. DOI: 10.1128/JCM.01280-15. eprint: https://jcm.asm.org/content/53/10/3334.full.pdf.

Van Boeckel, T. P. et al. (2015). "Global trends in antimicrobial use in food animals". In: *Proc Natl Acad Sci U S A* 112.18, pp. 5649–54. DOI: 10.1073/pnas.1503141112.

Ward, M. J. et al. (2014). "Time-Scaled Evolutionary Analysis of the Transmission and Antibiotic Resistance Dynamics of *Staphylococcus aureus* Clonal Complex 398". In: *Applied and Environmental Microbiology* 80.23. Ed. by C. A. Elkins, pp. 7275–7282. DOI: 10.1128/AEM.01777-14. eprint: https://aem.asm.org/content/80/23/7275.full.pdf.

White, David G. et al. (2001). "The Isolation of Antibiotic-Resistant *Salmonella* from Retail Ground Meats". In: *New England Journal of Medicine* 345.16. PMID: 11642230, pp. 1147–1154. DOI: 10.1056/NEJMoa010315. eprint: https://doi.org/10.1056/NEJMoa010315.

WHO (2014). *Antimicrobial resistance: global report on surveillance 2014*. WHO, Geneva, Switzerland.

— (2015). *WHO estimates of the global burden of foodborne diseases, 2007-2015*. WHO, Geneva, Switzerland.

Wong, Vanessa K. et al. (2015). "Phylogeographical analysis of the dominant multidrug-resistant H58 clade of *Salmonella* Typhi identifies inter- and intracontinental transmission events". In: *Nat Genet* 47.6, pp. 632–639. DOI: 10.1038/ng.3281.

Zankari, Ea et al. (2012). "Genotyping using whole-genome sequencing is a realistic alternative to surveillance based on phenotypic antimicrobial susceptibility testing". In: *Journal of Antimicrobial Chemotherapy* 68.4, pp. 771–777. DOI: 10.1093/jac/dks496. eprint: http://oup.prod.sis.lan/jac/article-pdf/68/4/771/2083079/dks496.pdf.

Zhang, S. et al. (2015). "*Salmonella* serotype determination utilizing high-throughput genome sequencing data". In: *J Clin Microbiol* 53.5, pp. 1685–92. DOI: 10.1128/JCM.00323-15.

Zhao, S. et al. (2016). "Whole-Genome Sequencing Analysis Accurately Predicts Antimicrobial Resistance Phenotypes in *Campylobacter* spp." In: *Appl Environ Microbiol* 82.2, pp. 459–466. DOI: 10.1128/AEM.02873-15.

CHAPTER 3

IDENTIFICATION OF NOVEL MOBILIZED COLISTIN RESISTANCE GENE *MCR-9* IN A MULTIDRUG-RESISTANT, COLISTIN-SUSCEPTIBLE *SALMONELLA ENTERICA* SEROTYPE TYPHIMURIUM ISOLATE[1]

[1]FROM CARROLL, LAURA M., AHMED GABALLA, CLAUDIA GULDIMANN, GENEVIEVE SULLIVAN, LORY O. HENDERSON, AND MARTIN WIEDMANN (2019). "IDENTIFICATION OF NOVEL MOBILIZED COLISTIN RESISTANCE GENE *MCR-9* IN A MULTIDRUG-RESISTANT, COLISTIN-SUSCEPTIBLE *SALMONELLA ENTERICA* SEROTYPE TYPHIMURIUM ISOLATE". IN: *MBIO* 10, PP. E00853-19. DOI: 10.1128/MBIO.00853-19.

3.1 Abstract

Mobilized colistin resistance (*mcr*) genes are plasmid-borne genes that confer resistance to colistin, an antibiotic used to treat severe bacterial infections. To date, eight known *mcr* homologues have been described (*mcr-1* to *-8*). Here, we describe *mcr-9*, a novel *mcr* homologue detected during routine *in silico* screening of sequenced *Salmonella* genomes for antimicrobial resistance genes. The amino acid sequence of *mcr-9*, detected in a multidrug-resistant (MDR) *Salmonella enterica* serotype Typhimurium (*S.* Typhimurium) strain isolated from a human patient in Washington State in 2010, most closely resembled *mcr-3*, aligning with 64.5% amino acid identity and 99.5% coverage using Translated Nucleotide BLAST (tblastn). The *S.* Typhimurium strain was tested for phenotypic resistance to colistin and was found to be sensitive at the 2-mg/liter European Committee on Antimicrobial Susceptibility Testing breakpoint under the tested conditions. *mcr-9* was cloned in colistin-susceptible *Escherichia coli* NEB5α under an IPTG (isopropyl-β-d-thiogalactopyranoside)-induced promoter to determine whether it was capable of conferring resistance to colistin when expressed in a heterologous host. Expression of *mcr-9* conferred resistance to colistin in *E. coli* NEB5α at 1, 2, and 2.5mg/liter colistin, albeit at a lower level than *mcr-3*. Pairwise comparisons of the predicted protein structures associated with all nine *mcr* homologues (Mcr-1 to -9) revealed that Mcr-9, Mcr-3, Mcr-4, and Mcr-7 share a high degree of similarity at the structural level. Our results indicate that *mcr-9* is capable of conferring phenotypic resistance to colistin in *Enterobacteriaceae* and should be immediately considered when monitoring plasmid-mediated colistin resistance.

IMPORTANCE: Colistin is a last-resort antibiotic that is used to treat se-

vere infections caused by MDR and extensively drug-resistant (XDR) bacteria. The World Health Organization (WHO) has designated colistin as a "highest priority critically important antimicrobial for human medicine" (WHO, Critically Important Antimicrobials for Human Medicine, 5th revision, 2017, https://www.who.int/foodsafety/publications/antimicrobials-fifth/en/), as it is often one of the only therapies available for treating serious bacterial infections in critically ill patients. Plasmid-borne *mcr* genes that confer resistance to colistin pose a threat to public health at an international scale, as they can be transmitted via horizontal gene transfer and have the potential to spread globally. Therefore, the establishment of a complete reference of *mcr* genes that can be used to screen for plasmid-mediated colistin resistance is essential for developing effective control strategies.

3.2 Observation

Until recently, bacterial resistance to colistin, a last-resort antibiotic reserved for treating severe infections, was thought to be acquired solely via chromosomal point mutations (Liu et al. 2016). However, in 2015, plasmid-mediated colistin resistance gene *mcr-1* was described in *Escherichia coli* (Liu et al. 2016). Mcr-1 is a phosphoethanolamine transferase that modifies cell membrane lipid A head groups with a phosphoethanolamine residue, reducing affinity to colistin (Anandan et al. 2017). Since then, seven additional *mcr* homologues (*mcr-2* to *-8*) have been identified in *Enterobacteriaceae* (Xavier et al. 2016; Yin et al. 2017; Carattoli, Villa, et al. 2017; Borowiak et al. 2017; AbuOun et al. 2017; Yang et al. 2018; Wang et al. 2018). Here, we report novel *mcr* homologue *mcr-9*, which was identified in a *Salmonella enterica* serotype Typhimurium (*S.* Typhimurium)

genome.

3.2.1 *In silico* identification of *mcr-9* in an MDR *S.* Typhimurium genome

MDR *S.* Typhimurium strain HUM_TYPH_WA_10_R9_3274 (NCBI RefSeq accession no. GCF_002091095.1) was isolated from a patient in Washington State in 2010 (Carroll, Wiedmann, et al. 2017). It had previously been tested for resistance to a panel of 12 antimicrobials that did not include colistin (Carroll, Wiedmann, et al. 2017). ABRicate version 0.8 (https://github.com/tseemann/abricate) identified 20 antimicrobial resistance (AMR) genes in the HUM_TYPH_WA_10_R9_3274 assembly using the ResFinder database (accessed 11 June 2018) (Zankari et al. 2012) and minimum identity and coverage thresholds of 75 and 50% (Carroll, Wiedmann, et al. 2017), respectively, none of which had been previously described to confer colistin resistance (see Table S1 in the supplemental material). Four plasmid replicons, including IncHI2 and IncHI2A, were detected with at least 80% identity and 60% coverage using ABRicate and PlasmidFinder (accessed 11 June 2018 [Table S1]) (Carattoli, Zankari, et al. 2014).

To detect *mcr-9* in the HUM_TYPH_WA_10_R9_3274 assembly, all colistin resistance-conferring nucleotide sequences available in ResFinder (52 sequences, accessed 22 January 2019 [see Table S2 in the supplemental material]) were translated into amino acid sequences using EMBOSS Transeq (reading frame 1 [https://www.ebi.ac.uk/Tools/st/emboss_transeq/]). The implementation of Translated Nucleotide BLAST (tblastn) (Camacho et al. 2009) in BTyper

version 2.3.2 (Carroll, Kovac, et al. 2017) selected *mcr-3.17* as the highest-scoring *mcr* allele, which aligned to *mcr-9* with 64.5% amino acid identity and 99.5% coverage (Table S1).

MUSCLE version 3.8.31 (Edgar 2004) was used to construct alignments of the amino acid sequence of *mcr-9* (NCBI protein accession no. WP_001572373.1) and the following: (i) the 52 *mcr* amino acid sequences from ResFinder (53 sequences [Table S2]), (ii) the top 100 hits produced when *mcr-9* was queried against NCBI's non-redundant protein (nr) database using the Protein BLAST (blastp) web server (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins [accessed 22 January 2019]; 152 sequences excluding *mcr-9*'s self-match [see Table S3 in the supplemental material]), and (iii) amino acid sequences of 61 putative phosphoethanolamine transferases used in other papers describing novel *mcr* genes (Yin et al. 2017; Carattoli, Villa, et al. 2017; Yang et al. 2018; Wang et al. 2018) (213 sequences [see Table S4 in the supplemental material]). For each alignment, RAxML version 8.2.12 (Stamatakis 2014) was used to construct a phylogeny using the PROTGAMMAAUTO method and 1,000 bootstrap replicates.

The amino acid sequence of *mcr-9* most closely resembled those of *mcr-3* and *mcr-7* (Figure 3.1; see Fig. S1 in the supplemental material). However, the *S.* Typhimurium isolate in which *mcr-9* was detected was not resistant to colistin at the > 2-mg/liter European Committee on Antimicrobial Susceptibility Testing (EUCAST [http://www.eucast.org]) breakpoint when a broth microdilution method was used to determine the colistin MIC (see Table S5 in the supplemental material).

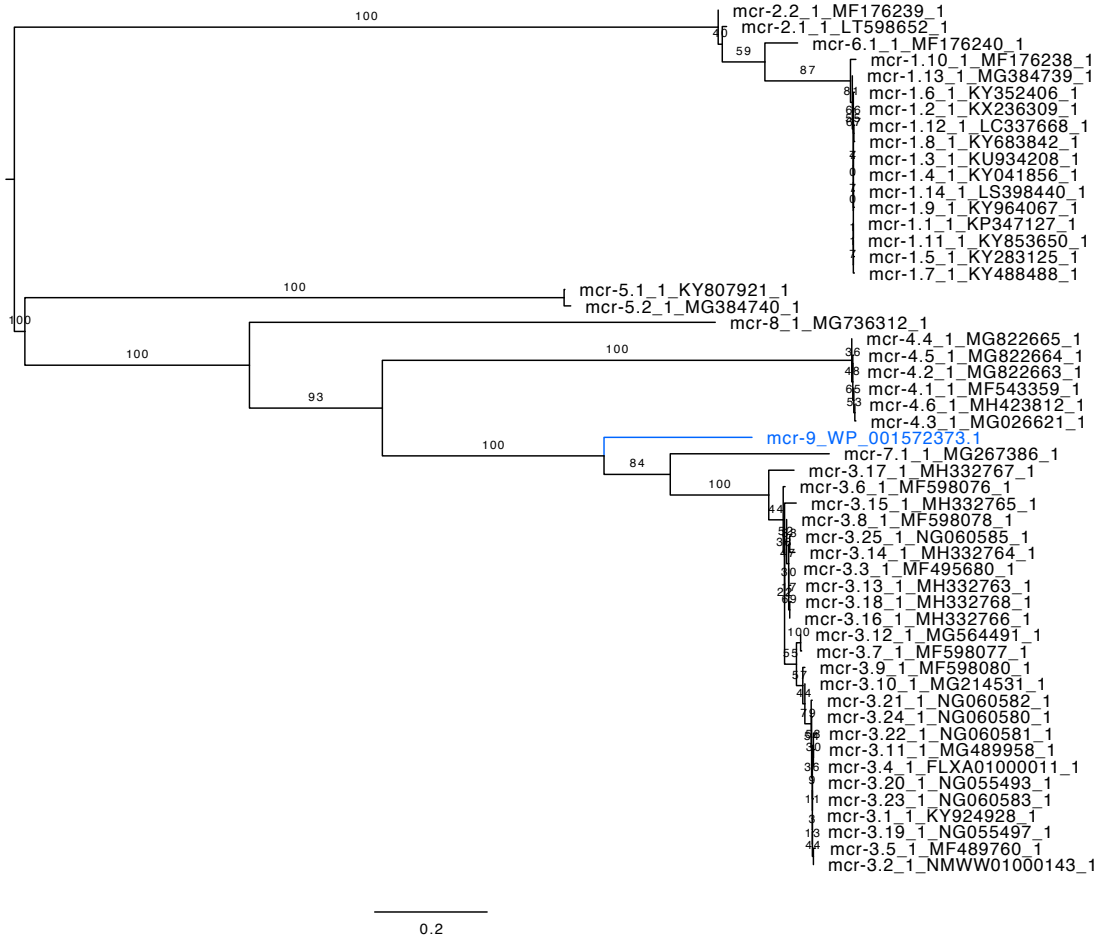


Figure 3.1: Comparison of *mcr-9* to all previously described *mcr* homologues, based on amino acid sequence. The maximum likelihood phylogeny was constructed using RAxML version 8.2.12 with the amino acid sequences of novel mobilized colistin resistance gene *mcr-9* (in blue) and all previously described *mcr* genes (*mcr-1* to *-8* [in black]). The phylogeny is rooted at the midpoint, with branch lengths reported in substitutions per site. Branch labels correspond to bootstrap support percentages out of 1,000 replicates.

3.2.2 *mcr-9* confers resistance to colistin when cloned into colistin-susceptible *E. coli* NEB5α

Coding regions of *mcr-9* and *mcr-3* were cloned under the control of an IPTG (isopropyl-β-d-thiogalactopyranoside)-induced SPAC/lacOid promoter and expressed in *E. coli* NEB5α (see Text S1 in the supplemental material). Colistin

killing assays (Figure 3.2; see Figure S2 in the supplemental material) were performed by incubating *E. coli* harboring the empty pLIV2 vector (negative control), pLIV2 with *mcr-3* (positive control), or pLIV2 with *mcr-9* with different concentrations of colistin (0, 1, 2, 2.5, and 5 mg/liter). *E. coli* cells harboring the empty vector failed to survive at all tested colistin concentrations > 0 mg/liter. While *mcr-3* expression conferred clinical levels of colistin resistance (i.e., beyond the 2-mg/liter EUCAST breakpoint) in *E. coli* at all tested concentrations, *mcr-9* expression conferred clinical resistance at 1, 2, and 2.5 mg/liter, but not 5 mg/liter of colistin (Figure 3.2; Figure S2).

3.2.3 Mcr-3, Mcr-4, Mcr-7, and Mcr-9 are highly similar at the structural level

Three-dimensional (3D) structural models of all nine Mcr homologues (Figure 3.3) based on EptA (Anandan et al. 2017) were constructed using the Phyre2 server (Kelley et al. 2015) and visualized using UCSF Chimera (Pettersen et al. 2004). Congruent with the phylogeny based on their amino acid sequences (Figure 3.1), comparisons of different Mcr protein models using Dali (Holm and Laakso 2016) revealed that Mcr-3, Mcr-4, Mcr-7, and Mcr-9 were closely related at the structural level (Figure 3.4).

Proteins encoded by *mcr-1* to *-9* revealed high levels of conservation for both the membrane-anchored domain and the soluble catalytic domain (Figure 3.3). Interestingly, analyses of structural models of the nine Mcr homologues using the ESPript 3 server (Robert and Gouet 2014) showed that both amino acids and structural elements were conserved on the C-terminal catalytic domain,

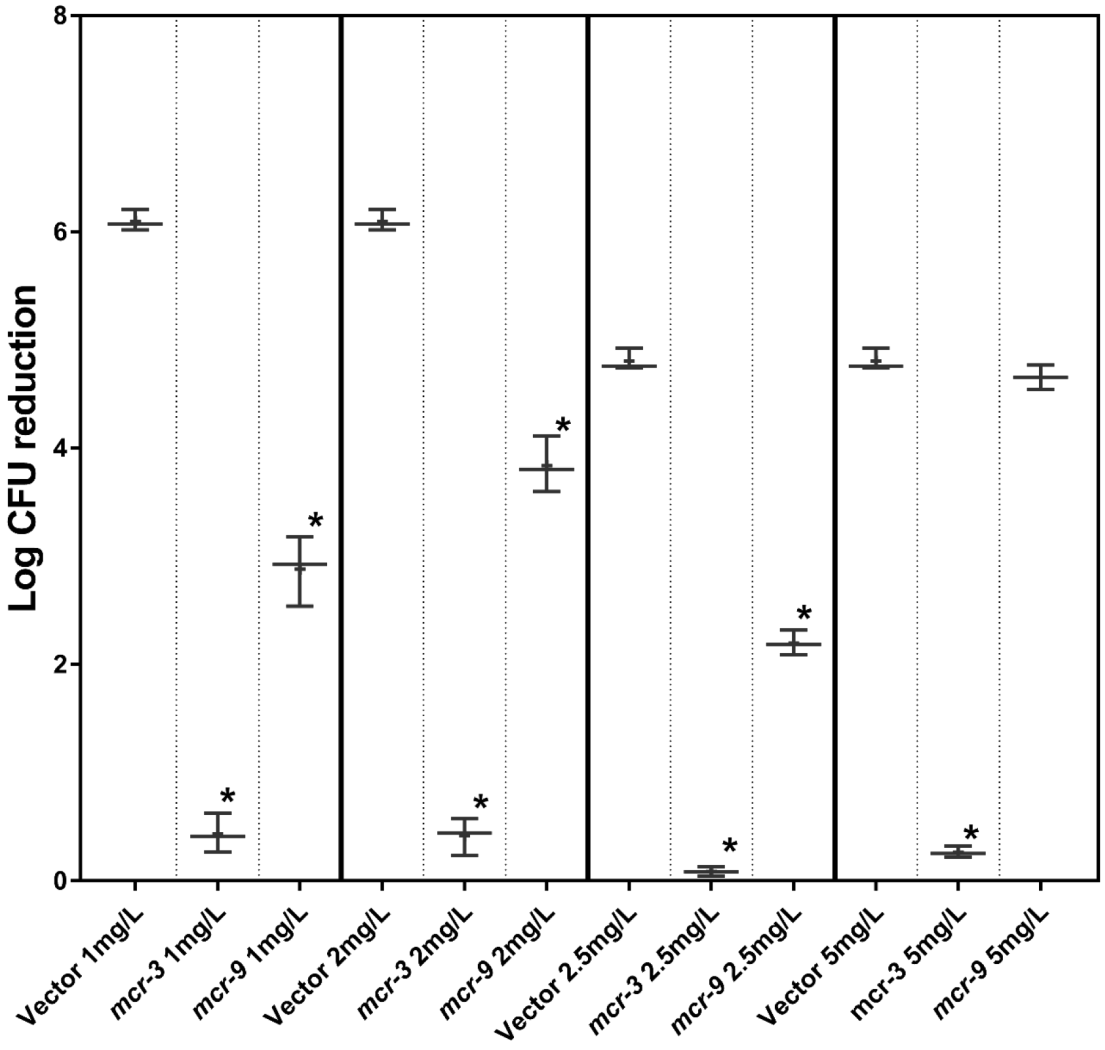


Figure 3.2: Colistin killing assay of *E. coli* NEB5α harboring a pLIV2 empty vector (negative control), *mcr-3* (positive control), or *mcr-9*, expressed under the control of the IPTG-controlled SPAC/lacOid promoter. Cells were grown in MH-II (Mueller-Hinton II) medium with IPTG to the mid-exponential phase. Colistin was added at concentrations of 0, 1, 2, 2.5, or 5 mg/liter, and the bacteria were incubated at 37°C for 1h. The samples were diluted in phosphate-buffered saline (PBS) and plated on LB agar plates for the determination of CFU. Log CFU reduction was calculated by comparing CFU after each treatment to CFU levels obtained at 0 mg/liter colistin, using three independent biological replicates. Asterisks denote significant differences compared to empty vector treatment ($P < 0.05$ by Student's t test relative to the concentration's respective negative control after a Bonferroni correction).

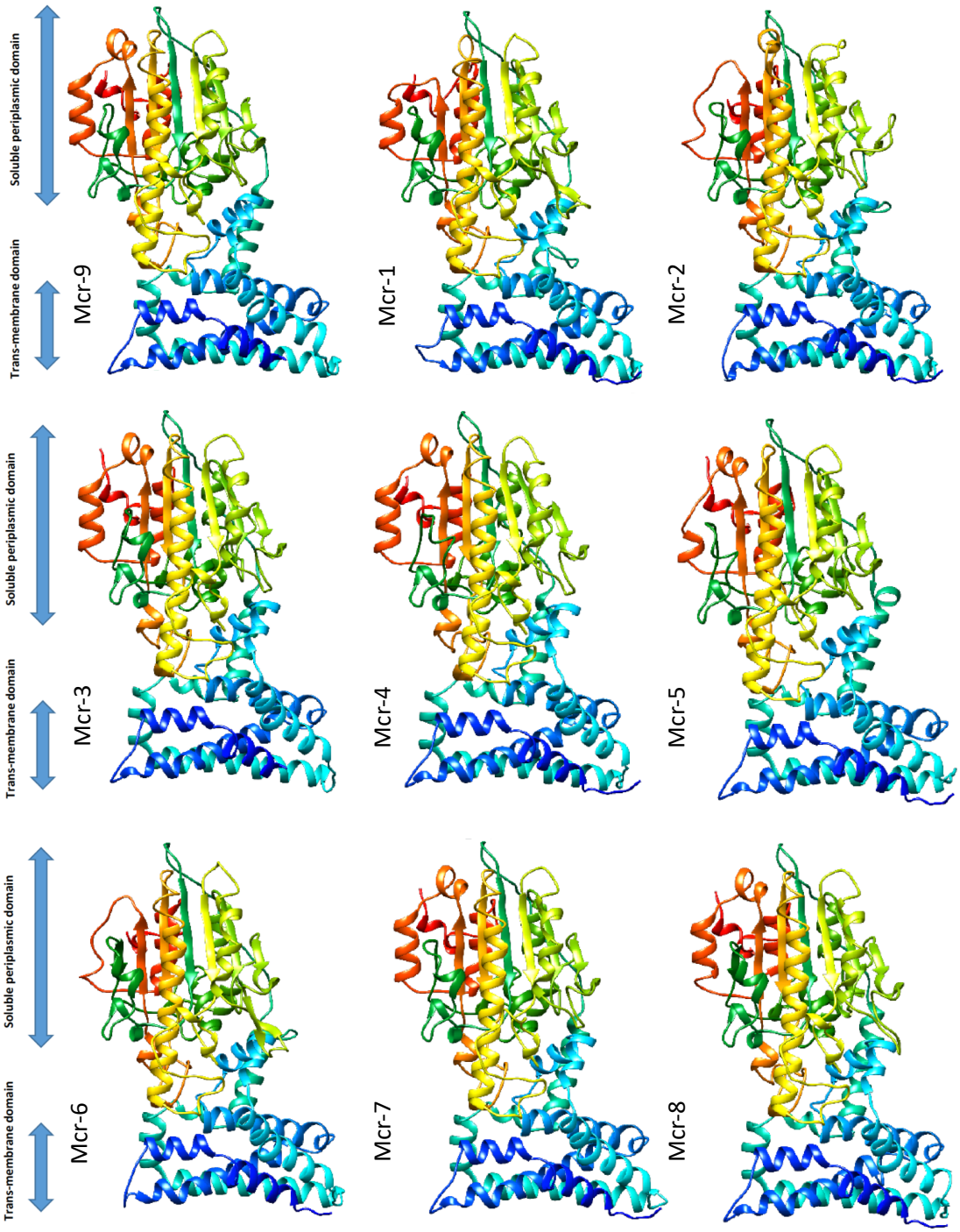


Figure 3.3: Structural models of all published Mcr proteins (Mcr-1 to -8) and Mcr-9, based on lipooligosaccharide phosphoethanolamine transferase EptA. Models were constructed using the Phyre2 server, and structures were viewed and edited using UCSF Chimera. Structural models show conservation of two EptA domains: transmembrane-anchored and soluble periplasmic domains.

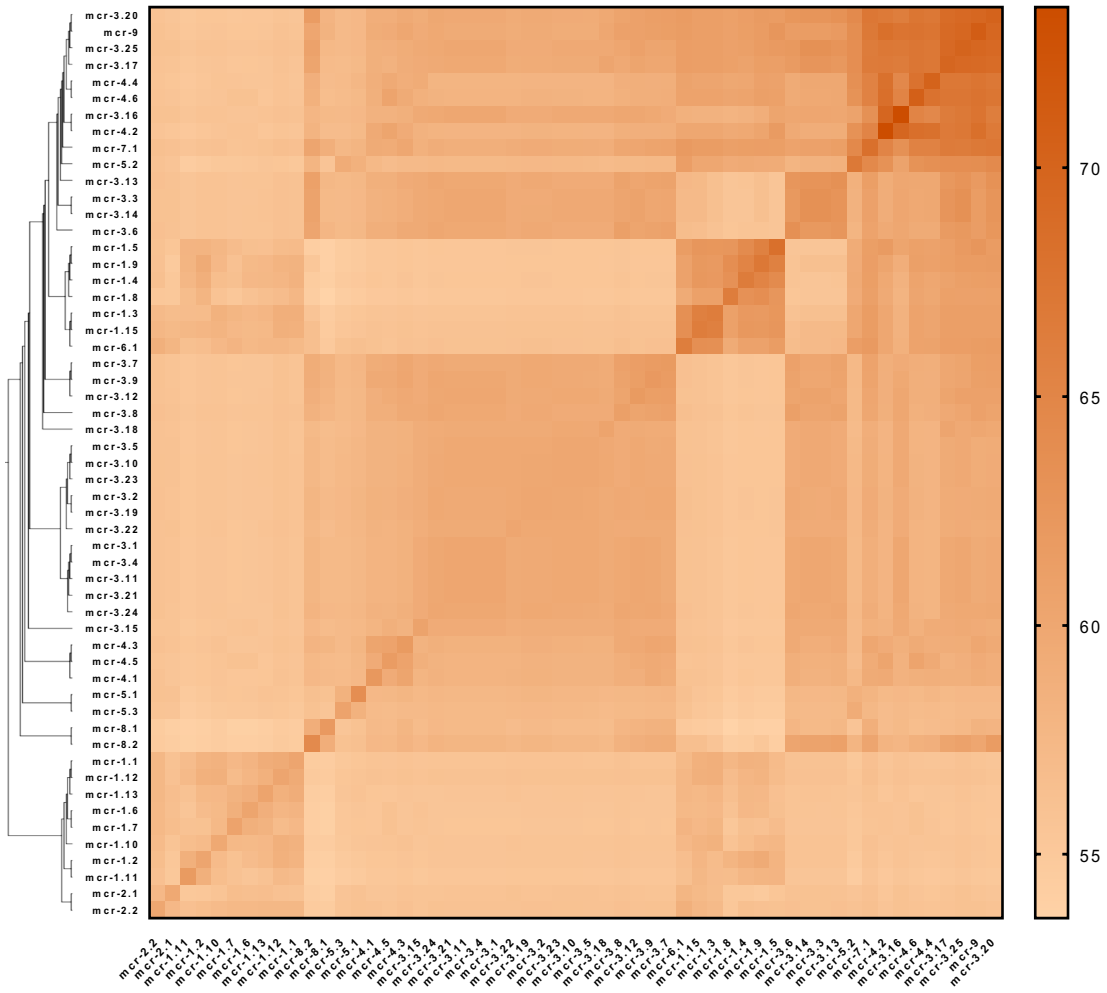


Figure 3.4: Similarity matrix (composed of Dali Z-scores) of all previously described Mcr groups (Mcr-1 to -8) and Mcr-9, based on protein structure. The Dali server was used to perform all-against-all comparisons of 3D structural models based on all *mcr* homologues (Figure 3.3); for this analysis, amino acid sequences of *mcr-5.3* and *mcr-8.2*, which were not available in ResFinder, were additionally included from the National Database of Antibiotic Resistant Organisms (NDARO).

while only structural elements were conserved on the membrane-anchored N-terminal domain (Figure 3.5).

3.2.4 Numerous genera of *Enterobacteriaceae* harbor *mcr-9* on IncHI2 plasmids.

blastp searches of *mcr-9* against NCBI's nr database revealed that *mcr-9* was present in multiple genera of *Enterobacteriaceae* (Table S3). The 10 highest-scoring hits in the nr database matched *mcr-9* with at least 99% amino acid identity (including *mcr-9* characterized here [Table S3 and Figure S1A]); the amino acid identities of the remaining hits with high query coverage (> 90%) dropped below 88% identity (Table S3 and Figure S1A). *mcr-9* was detected in 335 genomes linked to NCBI identical protein groups (IPGs) associated with the 10 highest-scoring protein accession numbers (accessed 23 January 2019 [see Tables S3 and S6 in the supplemental material]). Analysis of the *mcr-9* promoter region in 321 of these genomes (Text S1) showed conserved putative σ^{70} family-dependent -35 and -10 regions and an inverted repeat (Figure 3.6). The conserved DNA motif in the *mcr-9* promoter is likely a recognition sequence for a transcription regulator, suggesting that additional factors or induction/derepression conditions might be needed for full expression of wild-type *mcr-9*. Promoter variation (Huang et al. 2018) and testing conditions (Zhang et al. 2017; Gwozdzinski et al. 2018) have been shown to influence *mcr* expression and the colistin MIC, which may explain why the *S.* Typhimurium strain queried here was colistin susceptible under the tested conditions.

Of the 335 genomes in which *mcr-9* was detected, 65 had at least one plas-

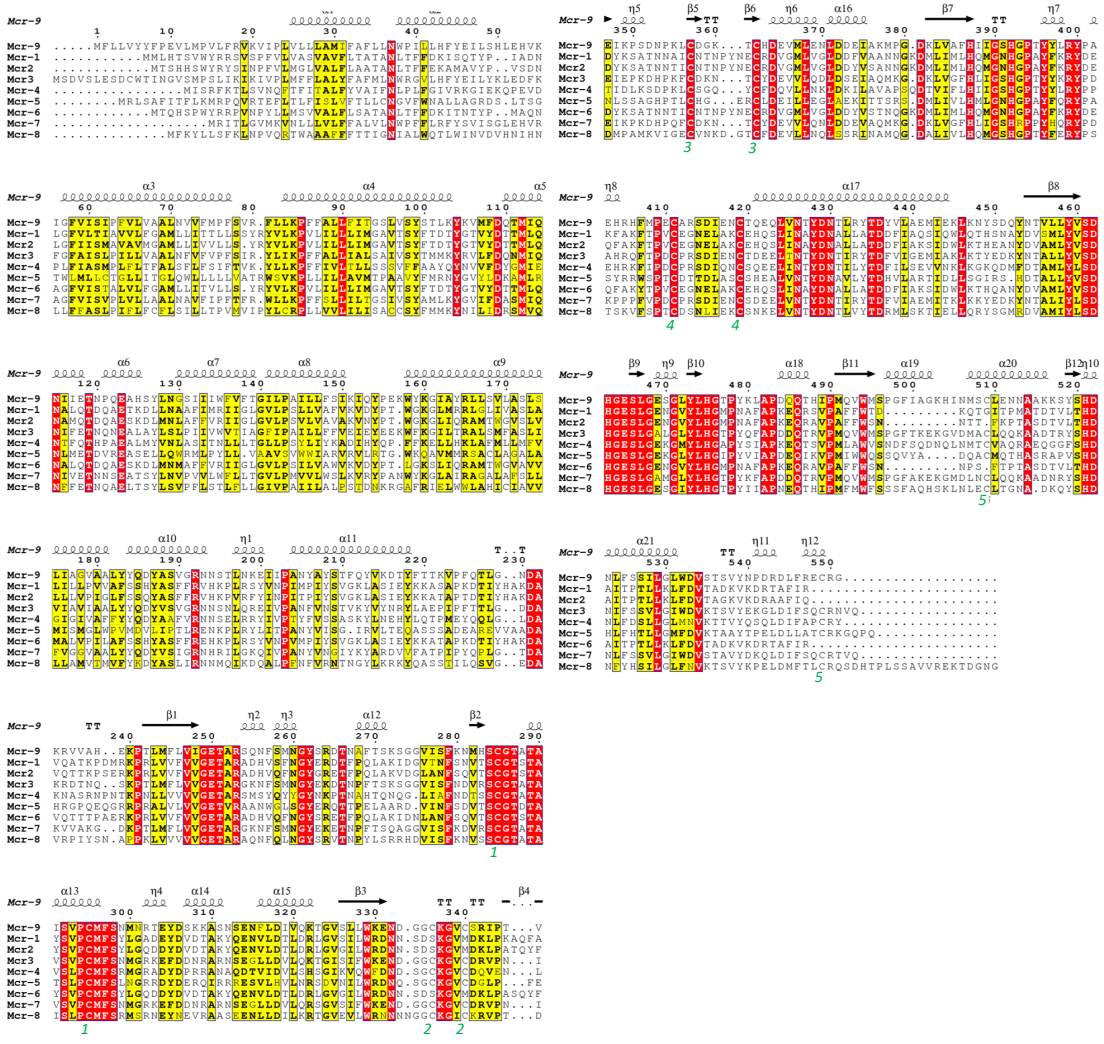


Figure 3.5: Location of Mcr-9 secondary structure elements within the alignment of Mcr amino acid sequences, constructed using the ESPript 3 server. The top track denotes Mcr-9 secondary structure elements (alpha helixes and beta sheets). Green digits below the alignment denote cysteine residues forming a disulfide bridge (e.g., 1 forms a bridge with 1, 2 with 2, etc.). Within the amino acid sequence alignment itself, a strict identity (i.e., identical amino acid residue at a site) is denoted by a red box and a white character. A yellow box around an amino acid residue denotes similarity across groups, where groups were defined using the default "all" specification in ESPript 3 (*ESPript 3 total score [TSc] > in-group threshold [ThIn]*), while a residue in boldface denotes similarity within a group (*ESPript 3 in-group score [ISc] > ThIn*).

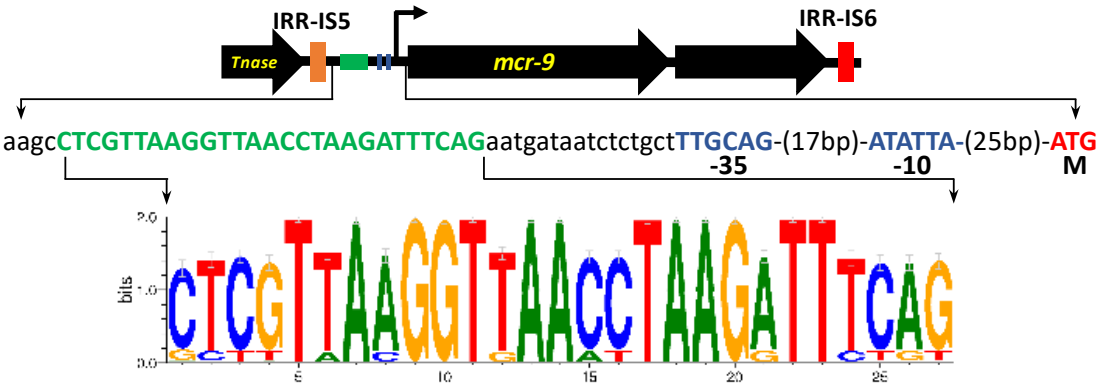


Figure 3.6: Organization of the *mcr-9* locus in *S.* Typhimurium. An unknown function cupin fold metalloprotein is encoded by the gene downstream of *mcr-9* (unlabeled black arrow). The *mcr-9* locus is flanked by two different terminal repeat sequences (IRR) from the IS5 (orange box) and IS6 (red box) families. The *mcr-9* upstream region contains highly conserved putative -35 and -10 σ^{70}-dependent promoter elements (blue boxes and blue text). Moreover, the *mcr-9* promoter region contains an inverted repeat motif (green box, green text, and sequence logo) that is conserved in more than 95% of 321 *mcr-9* genes, as shown by the sequence logo (constructed using WebLogo) (Crooks et al. 2004).

mid replicon (detected using ABRicate and PlasmidFinder as described above) present on the same contig as *mcr-9*; in 59 of these 65 genomes, IncHI2 and/or IncHI2A replicons were detected on the same contig as *mcr-9* (Table S6). In 32 of the 37 closed genomes in which it was detected, *mcr-9* was harbored on a plasmid (Table S6). These results indicate that *mcr-9* has the potential to reduce susceptibility to colistin, up to and beyond the EUCAST breakpoint, and can be found extrachromosomally in multiple species of *Enterobacteriaceae*, making it a relevant threat to public health. Future studies querying the plasmids that harbor *mcr-9* (e.g., transferability, stability, and copy number variation) will offer further insight into the potential role that *mcr-9* plays in the dissemination of colistin resistance worldwide.

3.2.5 Accession number(s) and supplemental material

The nucleotide and amino acid sequences of *mcr-9* are available under NCBI reference sequence accession no. NZ_NAAN01000063.1 (NCBI protein accession no. WP_001572373.1). Supplemental material is available at https://mbio.asm.org/content/10/3/e00853-19/figures-only.

3.3 Acknowledgments

This material is based on work supported by the National Science Foundation (NSF) Graduate Research Fellowship Program under grant no. DGE-1650441, with additional funding provided by an NSF Graduate Research Opportunities Worldwide (GROW) grant through a partnership with the Swiss National Science Foundation (SNF).

We thank Julie Siler (Cornell University) for providing colistin resistance testing materials.

3.4 References

AbuOun, M. et al. (2017). "*mcr-1* and *mcr-2* variant genes identified in *Moraxella* species isolated from pigs in Great Britain from 2014 to 2015". In: *J Antimicrob Chemother* 72.10, pp. 2745–2749. DOI: 10.1093/jac/dkx286.

Anandan, A. et al. (2017). "Structure of a lipid A phosphoethanolamine transferase suggests how conformational changes govern substrate binding". In: *Proc Natl Acad Sci U S A* 114.9, pp. 2218–2223. DOI: 10.1073/pnas.1612927114.

Borowiak, M. et al. (2017). "Identification of a novel transposon-associated phosphoethanolamine transferase gene, *mcr-5*, conferring colistin resistance in d-tartrate fermenting *Salmonella enterica* subsp. enterica serovar Paratyphi B". In: *J Antimicrob Chemother* 72.12, pp. 3317–3324. DOI: 10.1093/jac/dkx327.

Camacho, C. et al. (2009). "BLAST+: architecture and applications". In: *BMC Bioinformatics* 10, p. 421. DOI: 10.1186/1471-2105-10-421.

Carattoli, A., L. Villa, et al. (2017). "Novel plasmid-mediated colistin resistance *mcr-4* gene in *Salmonella* and *Escherichia coli*, Italy 2013, Spain and Belgium, 2015 to 2016". In: *Euro Surveill* 22.31. DOI: 10.2807/1560-7917.ES.2017.22.31.30589.

Carattoli, A., E. Zankari, et al. (2014). "*In silico* detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing". In: *Antimicrob Agents Chemother* 58.7, pp. 3895–903. DOI: 10.1128/AAC.02412-14.

Carroll, L. M., J. Kovac, R. A. Miller, and M. Wiedmann (2017). "Rapid, high-throughput identification of anthrax-causing and emetic *Bacillus cereus* group genome assemblies using BTyper, a computational tool for virulence-based classification of *Bacillus cereus* group isolates using nucleotide sequencing data". In: *Appl Environ Microbiol*. DOI: 10.1128/AEM.01096-17.

Carroll, L. M., M. Wiedmann, et al. (2017). "Whole-Genome Sequencing of Drug-Resistant *Salmonella enterica* Isolates from Dairy Cattle and Humans in New York and Washington States Reveals Source and Geographic Associations". In: *Appl Environ Microbiol* 83.12. DOI: 10.1128/AEM.00140-17.

Crooks, G. E., G. Hon, J. M. Chandonia, and S. E. Brenner (2004). "WebLogo: a sequence logo generator". In: *Genome Res* 14.6, pp. 1188–90. DOI: 10.1101/gr.849004.

Edgar, R. C. (2004). "MUSCLE: multiple sequence alignment with high accuracy and high throughput". In: *Nucleic Acids Res* 32.5, pp. 1792–7. DOI: 10.1093/nar/gkh340.

Gwozdzinski, K., S. Azarderakhsh, C. Imirzalioglu, L. Falgenhauer, and T. Chakraborty (2018). "An Improved Medium for Colistin Susceptibility Testing". In: *J Clin Microbiol* 56.5. DOI: 10.1128/JCM.01950-17.

Holm, L. and L. M. Laakso (2016). "Dali server update". In: *Nucleic Acids Res* 44.W1, W351–5. DOI: 10.1093/nar/gkw357.

Huang, B. et al. (2018). "Promoter Variation and Gene Expression of *mcr-1*-Harboring Plasmids in Clinical Isolates of *Escherichia coli* and *Klebsiella pneumoniae* from a Chinese Hospital". In: *Antimicrob Agents Chemother* 62.5. DOI: 10.1128/AAC.00018-18.

Kelley, L. A., S. Mezulis, C. M. Yates, M. N. Wass, and M. J. Sternberg (2015). "The Phyre2 web portal for protein modeling, prediction and analysis". In: *Nat Protoc* 10.6, pp. 845–58. DOI: 10.1038/nprot.2015.053.

Liu, Y. Y. et al. (2016). "Emergence of plasmid-mediated colistin resistance mechanism *MCR-1* in animals and human beings in China: a microbiological and molecular biological study". In: *Lancet Infect Dis* 16.2, pp. 161–8. DOI: 10.1016/S1473-3099(15)00424-7.

Pettersen, E. F. et al. (2004). "UCSF Chimera–a visualization system for exploratory research and analysis". In: *J Comput Chem* 25.13, pp. 1605–12. DOI: 10.1002/jcc.20084.

Robert, X. and P. Gouet (2014). "Deciphering key features in protein structures with the new ENDscript server". In: *Nucleic Acids Res* 42.Web Server issue, W320–4. DOI: 10.1093/nar/gku316.

Stamatakis, A. (2014). "RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies". In: *Bioinformatics* 30.9, pp. 1312–3. DOI: 10.1093/bioinformatics/btu033.

Wang, X. et al. (2018). "Emergence of a novel mobile colistin resistance gene, *mcr-8*, in NDM-producing *Klebsiella pneumoniae*". In: *Emerg Microbes Infect* 7.1, p. 122. DOI: 10.1038/s41426-018-0124-z.

Xavier, B. B. et al. (2016). "Identification of a novel plasmid-mediated colistin-resistance gene, *mcr-2*, in *Escherichia coli*, Belgium, June 2016". In: *Euro Surveill* 21.27. DOI: 10.2807/1560-7917.ES.2016.21.27.30280.

Yang, Y. Q., Y. X. Li, C. W. Lei, A. Y. Zhang, and H. N. Wang (2018). "Novel plasmid-mediated colistin resistance gene *mcr-7.1* in *Klebsiella pneumoniae*". In: *J Antimicrob Chemother*. DOI: 10.1093/jac/dky111.

Yin, W. et al. (2017). "Novel Plasmid-Mediated Colistin Resistance Gene *mcr-3* in *Escherichia coli*". In: *MBio* 8.3. DOI: 10.1128/mBio.00543-17.

Zankari, E. et al. (2012). "Identification of acquired antimicrobial resistance genes". In: *J Antimicrob Chemother* 67.11, pp. 2640–4. DOI: 10.1093/jac/dks261.

Zhang, H. et al. (2017). "Expression characteristics of the plasmid-borne *mcr-1* colistin resistance gene". In: *Oncotarget* 8.64, pp. 107596–107602. DOI: 10.18632/oncotarget.22538.

CHAPTER 4

RAPID, HIGH-THROUGHPUT IDENTIFICATION OF ANTHRAX-CAUSING AND EMETIC *BACILLUS CEREUS* GROUP GENOME ASSEMBLIES VIA BTYPER, A COMPUTATIONAL TOOL FOR VIRULENCE-BASED CLASSIFICATION OF *BACILLUS CEREUS* GROUP ISOLATES BY USING NUCLEOTIDE SEQUENCING DATA[1]

[1]FROM CARROLL, LAURA M., JASNA KOVAC, RACHEL A. MILLER, AND MARTIN WIEDMANN (2017). "RAPID, HIGH-THROUGHPUT IDENTIFICATION OF ANTHRAX-CAUSING AND EMETIC *BACILLUS CEREUS* GROUP GENOME ASSEMBLIES VIA BTYPER, A COMPUTATIONAL TOOL FOR VIRULENCE-BASED CLASSIFICATION OF *BACILLUS CEREUS* GROUP ISOLATES BY USING NUCLEOTIDE SEQUENCING DATA". IN: *APPLIED AND ENVIRONMENTAL MICROBIOLOGY* 83, PP. E01096-17. DOI: 10.1128/AEM.01096-17.

4.1 Abstract

The *Bacillus cereus* group comprises nine species, several of which are pathogenic. Differentiating between isolates that may cause disease and those that do not is a matter of public health and economic importance, but it can be particularly challenging due to the high genomic similarity within the group. To this end, we have developed BTyper, a computational tool that employs a combination of (i) virulence gene-based typing, (ii) multilocus sequence typing (MLST), (iii) *panC* clade typing, and (iv) *rpoB* allelic typing to rapidly classify *B. cereus* group isolates using nucleotide sequencing data. BTyper was applied to a set of 662 *B. cereus* group genome assemblies to (i) identify anthrax-associated genes in non-*B. anthracis* members of the *B. cereus* group, and (ii) identify assemblies from *B. cereus* group strains with emetic potential. With BTyper, the anthrax toxin genes *cya*, *lef*, and *pagA* were detected in 8 genomes classified by the NCBI as *B. cereus* that clustered into two distinct groups using *k*-medoids clustering, while either the *B. anthracis* poly-γ-d-glutamate capsule biosynthesis genes *capABCDE* or the hyaluronic acid capsule *hasA* gene was detected in an additional 16 assemblies classified as either *B. cereus* or *Bacillus thuringiensis* isolated from clinical, environmental, and food sources. The emetic toxin genes *cesABCD* were detected in 24 assemblies belonging to *panC* clades III and VI that had been isolated from food, clinical, and environmental settings. The command line version of BTyper is available at https://github.com/lmc297/BTyper. In addition, BMiner, a companion application for analyzing multiple BTyper output files in aggregate, can be found at https://github.com/lmc297/BMiner.

IMPORTANCE: *Bacillus cereus* is a foodborne pathogen that is estimated to

cause tens of thousands of illnesses each year in the United States alone. Even with molecular methods, it can be difficult to distinguish nonpathogenic *B. cereus* group isolates from their pathogenic counterparts, including the human pathogen *Bacillus anthracis*, which is responsible for anthrax, as well as the insect pathogen *B. thuringiensis*. By using the variety of typing schemes employed by BTyper, users can rapidly classify, characterize, and assess the virulence potential of any isolate using its nucleotide sequencing data.

4.2 Introduction

The *Bacillus cereus* group, also known as *Bacillus cereus sensu lato* (*s.l.*), consists of nine closely related bacterial species: *B. anthracis* (Logan 2015), *B. cereus sensu stricto* (*s.s.*), *B. cytotoxicus* (Guinebretiere, Auger, et al. 2013), *B. mycoides* (Lechner et al. 1998), *B. pseudomycoides* (Nakamura 1998), *B. thuringiensis*, *B. toyonensis* (G. Jimenez et al. 2013), *B. weihenstephanensis* (Lechner et al. 1998), and *B. wiedmannii* (Miller, Beno, et al. 2016). The pathogenic potentials of members of the *B. cereus* group vary widely; while some isolates are capable of causing anthrax or anthrax-like disease (CDC n.d.), foodborne illness (Stenfors Arnesen, Fagerlund, and Granum 2008), or food spoilage issues (Lucking et al. 2013; Doll, Scherer, and Wenning 2017; Ivy et al. 2012), others are used in industrial settings as probiotics (G. Jimenez et al. 2013; Hong, Le Hong Duc, and Cutting 2005; Guillermo Jimenez et al. 2013; Zhu et al. 2016), insecticides and pest control agents (Jouzani, Valijanian, and Sharafi 2017), agents in environmental pollutant bioremediation (Jouzani, Valijanian, and Sharafi 2017; Aceves-Diez, Estrada-Castaneda, and Castaneda-Sandoval 2015; Dash, Mangwani, and Das 2014), plant growth promoters (Jouzani, Valijanian, and Sharafi 2017; Ar-

mada et al. 2015), and even as producers of bacteriocins (Wang et al. 2014; Lee, Churey, and Worobo 2009) or parasporins with anticancer activities (Jouzani, Valijanian, and Sharafi 2017; Ohba, Mizuki, and Uemori 2009; Ammons et al. 2016). As the industrial and agricultural applications of these microorganisms expand, differentiating between isolates that can cause anthrax or gastrointestinal illness and those that can be used as beneficial microbes in industrial or agricultural settings becomes critical. Relying strictly on taxonomic classification at the species level can lead not only to isolate misclassification, but also to an inaccurate assessment of a given isolate's virulence potential. There have been numerous cases in which probiotics containing *B. cereus* group isolates sold for human and/or animal consumption were found to possess strains capable of producing toxins Nhe and/or Hbl (Hong, Le Hong Duc, and Cutting 2005; Zhu et al. 2016; Le H. Duc et al. 2004), or the species they contained were incorrectly identified (Hong, Le Hong Duc, and Cutting 2005; Zhu et al. 2016; Huys et al. 2013). Additionally, *B. thuringiensis*, a biopesticide, can possess *B. cereus s.s.* toxin genes and potentially infect humans via the food chain (Rosenquist et al. 2005), a notable example being a foodborne outbreak associated with salad that was potentially caused by *B. thuringiensis* serovar aizawai that had been sprayed on a produce field (EFSA 2016).

Differentiating between pathogenic and nonpathogenic *B. cereus* group isolates is a matter of public health and economic importance but can be a challenging task. Phenotypic and biochemical methods (Tallent et al. 2012), as well as many commonly used molecular methods, such as 16S rRNA gene sequencing, may not have sufficient discriminatory power to differentiate between members of the *B. cereus* group (Liu et al. 2015a; Fox, Wisotzkey, and Jurtshuk 1992). In addition, the ability of a particular *B. cereus* group isolate to cause disease in

humans is not species dependent, and taxonomic classification can often be a poor predictor of an isolate's virulence potential (Kovac et al. 2016); for example, genes encoding diarrheal toxins have been found in *B. cereus*, *B. mycoides*, *B. pseudomycoides*, *B. thuringiensis*, and *B. weihenstephanensis* (Kovac et al. 2016; Izabela Swiecicka, Van der Auwera, and Mahillon 2006; Pruss et al. 1999). For these reasons, better tools are needed to classify *B. cereus* isolates, from both taxonomical and food safety risk perspectives (Ehling-Schulz and Messelhausser 2013).

A number of genetic loci have been proposed as markers that can be used to taxonomically classify and/or differentiate between pathogenic and non-pathogenic *B. cereus* group isolates at greater resolution than phenotypic methods and 16S rRNA gene sequencing (Kovac et al. 2016). Some examples of taxonomic markers include the housekeeping gene *rpoB* (Miller, Beno, et al. 2016; Kovac et al. 2016; Caamano-Antelo et al. 2015; Kwan Soo Ko et al. 2004; K. S. Ko et al. 2003; Martinez, Stratton, and Bianchini 2017; Miller, Kent, et al. 2015), the pantoate-beta-alanine ligase gene *panC* (Guinebretiere, Thompson, et al. 2008; Guinebretiere, Velge, et al. 2010; Warda et al. 2016; Schmid et al. 2016; Sorokin et al. 2006), and multiple loci used in a 7-gene multilocus sequence typing (MLST) scheme (i.e., *glp*, *gmk*, *ilv*, *pta*, *pur*, *pyc*, and *tpi*) (Kovac et al. 2016; Yang, Yu, et al. 2017; Yang, Gu, et al. 2016; Drewnowska and Izabela Swiecicka 2013; Tourasse et al. 2011; A. R. Hoffmaster et al. 2008; Cardazzo et al. 2008) (https://pubmlst.org/bcereus/). Each of these methods alone provides greater resolution than its predecessors, and the methods may be implemented in combination with each other and/or with phenotypic methods (Kovac et al. 2016; Ehling-Schulz and Messelhausser 2013; Guinebretiere, Velge, et al. 2010; Cardazzo et al. 2008).

The presence and absence of virulence and toxin genes have also served as indicators in a method by which *B. cereus* group isolates can be classified as pathogenic or nonpathogenic (Liu et al. 2015b; Kovac et al. 2016; Bohm et al. 2015). These methods are beneficial from a clinical perspective, as genes associated with many medically relevant phenotypes are plasmid carried (Klee et al. 2010), including anthrax toxin and capsule genes (Zwick et al. 2012), and *ces* genes, which encode cereulide synthetase (Hoton et al. 2009). This can be contrasted with the fact that many genes that encode phenotypic traits used to distinguish members of the *B. cereus* group using biochemical and microbiological tests are contained on the chromosome (motility, hemolysis, etc.) (Klee et al. 2010). As a result, a disease phenotype, such as the ability to cause anthrax-like symptoms in a particular host (Zwick et al. 2012), may not be confined to a single *B. cereus* group species, making species-level taxonomy a poor indicator of an isolate's pathogenic potential.

Molecular typing methods using housekeeping and virulence genes found in members of the *B. cereus* group have been essential for classifying isolates from both a taxonomical and a public health perspective. However, as whole-genome sequencing (WGS) becomes cheaper, faster, and more accessible, the ability to perform molecular typing methods *in silico* becomes even more attractive. With the goal of creating a readily accessible open-source pipeline that can be easily used by *B. cereus* researchers and public health officials, we have created BTyper, a computational tool to perform (i) virulence gene detection, (ii) MLST, (iii) *panC* clade typing, and (iv) *rpoB* allelic typing using *B. cereus* group nucleotide sequencing data in either FASTA, SRA, or gzipped FASTQ format. Additionally, we applied BTyper and BMiner, a companion application for analyzing BTyper's output files in aggregate, to a set of 662 *B. cereus* group

genome assemblies, with the goal of identifying (i) anthrax-associated genes in non-*anthracis Bacillus* members of the *B. cereus* group, and (ii) assemblies from *B. cereus* group strains with emetic potential.

4.3 Materials and Methods

4.3.1 Database construction

To construct a virulence gene database specific to *B. cereus* group isolates, amino acid sequences from a total of 36 virulence genes (see Table S1 in the supplemental material) were collected from the National Center for Biotechnology Information (NCBI) (https://www.ncbi.nlm.nih.gov/). For an MLST database, the 7-gene MLST database for *Bacillus cereus* was downloaded from PubMLST (https://pubmlst.org/bcereus/). For *panC* typing, chromosomes of 45 *B. cereus* group strains were downloaded from the NCBI database (Table S2). *panC* genes were extracted from each strain using nucleotide BLAST (BLASTn) (Camacho et al. 2009) and the *panC* genes of various *B. cereus* group type strains, and the online tool available at https://tools.symprevius.org/Bcereus/english.php was used to ensure that at least one representative from each of the seven *panC* clades was present in the collection (Guinebretiere, Velge, et al. 2010) (Table S2). For *rpoB* allelic typing, the *rpoB* allelic type database created and curated by Cornell University's Food Safety Lab and Milk Quality Improvement Program (CUFSL/MQIP; Ithaca, NY) was used. While 16S rRNA gene typing is not performed by default (see "Construction of BTyper tool," below), 16S rRNA gene typing can be performed using reference 16S rRNA gene sequences from nine

different *B. cereus* group type strain genomes. To obtain these sequences, the 16S rRNA gene sequence from a cultured *B. cereus* type strain was downloaded from the Ribosomal Database Project (RDP) (Cole et al. 2014) and used in conjunction with BLASTn (Camacho et al. 2009) to extract 16S rRNA gene genes from each of nine different *B. cereus* group species type strain genomes (Table S3). All database files can be downloaded from https://github.com/lmc297/BTyper.

4.3.2 Construction of BTyper tool

BTyper was created with the following dependencies: Python version 2.7 (https://www.python.org/), Biopython version 1.6.8 (Cock et al. 2009), BLAST version 2.4.0 (Camacho et al. 2009), SPAdes version 3.9.0 (Bankevich et al. 2012), and SRA toolkit version 2.8.0 (Kodama et al. 2012; Leinonen et al. 2011). The whole-genome sequences of 22 previously characterized *B. cereus* group isolates (Kovac et al. 2016) were downloaded from the NCBI and used as a training set to optimize parameters (referred to here as the "training set"; Table S4). For virulence gene detection using translated nucleotide BLAST (tBLASTn) (Camacho et al. 2009), default minimum coverage and minimum identity thresholds of 70 and 50%, were chosen, respectively, as they correlated highly with previously published PCR results (Kovac et al. 2016), and the allele with the highest corresponding bit score was reported. For MLST, *rpoB* allelic typing, and *panC* clade typing, the highest-scoring allele in the respective database was selected using its associated BLAST bit score, with no minimum threshold applied (Figure 4.1). Virulence gene detection, MLST, *rpoB* allelic typing, and *panC* clade typing methods were chosen to be performed by default, as these methods are valuable for their discriminatory power (Kovac et al. 2016). 16S rRNA gene

typing, although not performed by default due to its inability to discriminate between phylogenetic clades and species (Caamano-Antelo et al. 2015; Rossi-Tamisier et al. 2015; Chen and Tsen 2002), was added as an option as well, as many users may be interested in this locus. For this method, the highest-scoring 16S rRNA gene of the nine type strain 16S rRNA genes was selected using its BLAST bit score, with no minimum threshold applied.

4.3.3 PCR detection of virulence genes

To assess the accuracy of BTyper's *in silico* virulence gene detection, each of the 24 isolates in the validation set was screened for eight virulence genes (*hblA*, *hblC*, *hblD*, *nheA*, *nheB*, *nheC*, *cytK*, and *entFM*) using PCR. Bacterial DNA used as the template in PCRs was extracted by inoculating single colonies into 100 μl of sterile water; lysates were then heated at 95°C for 10 min in a thermocycler. For PCRs, 1 μl of dirty lysate was added to a master mix containing sterile water, 2x GoTaq Green master mix (Promega, Madison, WI), and primers at a concentration of 0.4 μM each (Table S5). The PCRs included an initial denaturation time of 3 min at 94°C, followed by 30 cycles of amplification; each cycle consisted of denaturation at 94°C for 30 s, annealing (see Table S5 for annealing temperatures) for 30 s, and elongation for 1 min at 72°C, with a final extension at 72°C for 7 min. PCR products were electrophoresed in 1% agarose gels, followed by ethidium bromide staining to confirm specific amplification. For isolates that did not yield a PCR amplicon for a given gene, the PCR was repeated at least once in order to confirm the negative PCR result.

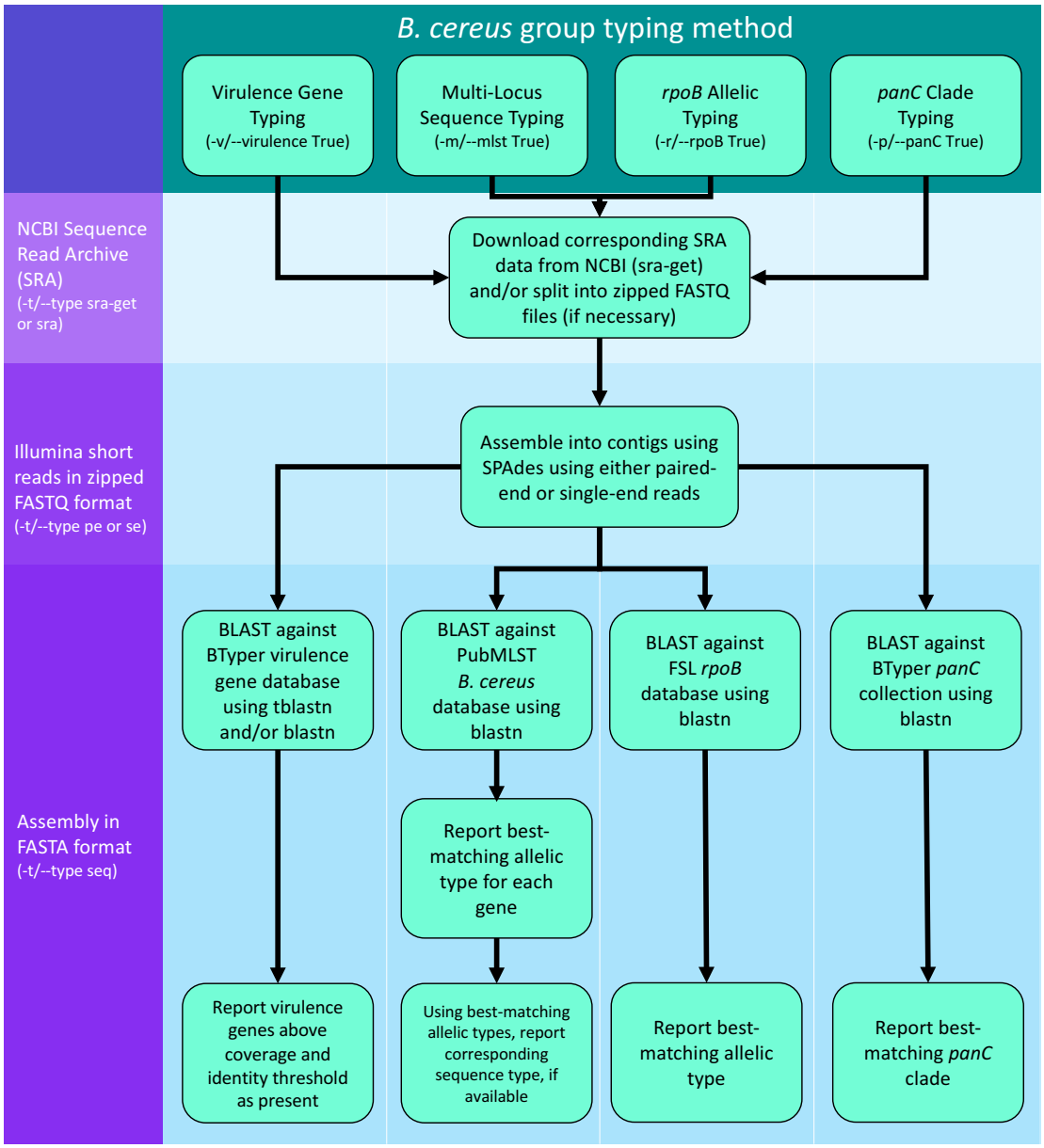


Figure 4.1: BTyper command line workflow for various types of data and default typing methods. Input datum type is listed in the left margin, while typing methods are listed at the top of the chart. Command line parameters associated with a particular typing method are shown in parentheses. FSL, Food Safety Lab.

4.3.4 MLST

Multilocus sequence typing (MLST) was performed for all 24 isolates in the validation set using a 7-housekeeping-gene scheme available through the PubMLST website (https://pubmlst.org/bcereus/). The PCRs consisted of 1 μl of dirty lysate as the DNA template added to a master mix containing sterile water, 2x GoTaq Green master mix (Promega), and primers at a final concentration of 0.4 μM each. The PCR cycles included an initial denaturation (3 min at 94°C), followed by 20 cycles of denaturation (94°C for 30 s), annealing for 30 s with a touchdown scheme (annealing temperatures that decrease by 0.5°C per cycle, starting with 55°C and reaching 45°C at the last cycle), and elongation at 72°C for 45 s. The 20 cycles of touchdown PCR were followed by an additional 20 cycles using an annealing temperature of 45°C. A final extension at 72°C for 5 min was included at the end of the 40 cycles. After amplification, the PCR products were sequenced at the Biotechnology Resource Center (BRC; Cornell University, Ithaca, NY), and ATs and sequence types (STs; based on all 7 genes) were assigned using the PubMLST website. All isolates were submitted to the *B. cereus* PubMLST database (Kovac et al. 2016).

4.3.5 *rpoB* allelic typing

A 632-nucleotide (nt) internal sequence of *rpoB*, encoding the β-subunit of the RNA polymerase, was used for assigning *rpoB* allelic types (ATs), as described previously (Ivy et al. 2012). The sequences of all *rpoB* ATs are available in the Food Microbe Tracker database (Vangay et al. 2013).

4.3.6 Validation of BTyper using additional *B. cereus* group whole-genome sequences

The genomes of 24 additional *B. cereus* group isolates were sequenced and assembled according to Miller et al. (referred to here as the "validation set"; Table S6) (Miller, Beno, et al. 2016). BTyper was used to perform virulence gene detection, MLST, *rpoB* allelic typing, and *panC* clade typing on each draft genome using the chosen default settings (see "Construction of BTyper tool", above). The same analyses were performed using the Illumina paired-end reads associated with each isolate, again using BTyper's default settings. To assess the accuracy of the *panC* clades assigned by BTyper, clade assignments provided by BTyper were compared to the isolates' whole-genome sequence clades provided by Kovac et al. (Kovac et al. 2016) and Miller et al. (Miller, Jian, et al. 2018) for the training and validation sets, respectively. A current version of the command line tool, as well as the curated virulence gene and *rpoB* allelic type databases, can be found at https://github.com/lmc297/BTyper. A link to a Web-based version of BTyper will also be made available at https://github.com/lmc297/BTyper at a later time.

4.3.7 Construction of BMiner companion application

BMiner, a companion application for parsing, viewing, and analyzing multiple BTyper files in aggregate, was created with the following dependencies: R version 3.3.2 (R Core Team 2016) and R packages shiny version 1.01 (Chang et al. 2017), ggplot2 version 2.2.1 (Wickham 2009), readr version 1.1.0 (Wickham, Hester, and Francois 2017), stringr version 1.2.0 (Wickham

2017), vegan version 2.4-2 (Oksanen et al. 2017), plyr version 1.8.4 (Wickham 2011), dplyr version 0.5.0 (Wickham, Francois, et al. 2016), cluster version 2.0.6 (Maechler et al. 2017), ggrepel version 0.6.5 (Slowikowski 2016), and magrittr version 1.5 (Bache and Wickham 2014). BMiner is freely available at https://github.com/lmc297/BMiner.

4.3.8 Application of BTyper and BMiner to whole-genome sequencing data

The latest assembly versions for all (n = 651) *B. cereus* group genome assemblies available in GenBank were downloaded on 6 April 2017. Genome assemblies were assigned to one of nine taxa according to their GenBank classification: *B. anthracis* (n = 157), *B. cereus s.s.* (n = 343), *B. cytotoxicus* (n = 2), *B. mycoides* (n = 19), *B. pseudomycoides* (n = 2), *B. thuringiensis* (n = 93), *B. toyonensis* (n = 3), *B. weihenstephanensis* (n = 21), and *B. wiedmannii* (n = 11). BTyper was used to perform virulence gene detection, MLST, *rpoB* allelic typing, and *panC* clade typing on all 651 isolates, as well as an additional 11 isolates that were part of the validation set but did not have assemblies in the NCBI database at the time (total number of *B. cereus* group genomes, 662). All available metadata associated with each assembly's BioSample were downloaded from the NCBI (Barrett et al. 2012). Data mining using BTyper results from all 662 *B. cereus* group assemblies was conducted using BMiner. The final results files for all 662 *B. cereus* group genome assemblies, as well as the associated metadata, can be found at https://github.com/lmc297/BTyper.

4.3.9 *Post hoc* statistical analyses

Post hoc statistical analyses were conducted in R version 3.3.2 (R Core Team 2016). Fisher's exact test was used to test for associations between virulence genes and *panC*-based phylogenetic clades using the fisher.test function in R's stats package (Table S7). Phylogenetic clades I and VII were excluded from this analysis, due to both being underrepresented among *B. cereus* group genomes in the NCBI database (12 and 2 isolates, respectively), while rare and common virulence genes present in fewer than 20 and more than $n - 20$ assemblies (where n corresponds to the total number of assemblies being tested), respectively, were also excluded. A Bonferroni correction was used to correct for multiple comparisons. To find members of the *B. cereus* group that clustered with *B. anthracis* isolates based on their virulence gene presence-absence profiles, as well as to assess within-species virulence heterogeneity, *k*-medoids clustering was performed using the clara function in R's cluster package (Maechler et al. 2017) and a Euclidean distance metric. To find an optimum value for k, k-medoids clustering was performed for each value of k for $2 \leq k \leq (n - 1)$, where n is 662, the total number of assembled genomes. A k value of 31 was selected, as it corresponded to the largest average silhouette width.

4.4 Results

4.4.1 Construction and validation of BTyper using *in vitro* methods

BTyper was used to perform *in silico* (i) virulence gene detection, (ii) MLST, (iii) *panC* clade typing, and (iv) *rpoB* allelic typing using the default settings described in Materials and Methods. Both assembled genomes and Illumina paired-end reads from 46 *B. cereus* group genomes were used (Figure 4.1). BTyper was successfully able to predict *rpoB* allelic types and whole-genome phylogenetic clade using *panC* for all *B. cereus* group genomes tested (n = 46; Table 4.1). For *in silico* MLST, it was successful at predicting the sequence type in all but one isolate (45 out of 46; Table 4.1); isolate FSL M8-0091 was the only isolate for which *in silico* prediction of sequence type did not match the sequence type obtained by Sanger sequencing. For this isolate, the only allele that differed between the two methods was the *tpi* allele: Sanger sequencing yielded a *tpi* allelic type of 20, while BTyper's *in silico* prediction was *tpi* allelic type 175, which was a perfect match and differed from *tpi* 20 by a single nucleotide at position 284. However, SRST2 (Inouye et al. 2014) also obtained a *tpi* allelic type of 175, making it likely that (i) the colony selected to undergo WGS had a different *tpi* allele than the colony selected to undergo Sanger sequencing, or (ii) there was an error in either WGS or Sanger sequencing.

For virulence gene detection, the results obtained from BTyper matched the PCR results for eight selected virulence genes in over 89% of all isolates (n = 46; Table 4.1). This resulted in an overall sensitivity and specificity of 99.0% and

Table 4.1: Percentage of isolates in which BTyper correctly identified the presence/absence of eight virulence genes, MLST, *rpoB* AT, and *panC* clade

Data set	Virulence gene (%)[a]										
	hblA	*hblC*	*hblD*	*nheA*	*nheB*	*nheC*	*cytK*	*entFM*	*MLST ST (%)[b]*	*rpoB AT (%)[c]*	*panC clade (%)[d]*
Training (n = 22)											
Assemblies	100	100	100	100	95.5	100	90.9	95.5	100	100	100
PE reads[e]	100	90.9	100	90.9	95.5	95.5	90.9	95.5	100	100	100
Validation (n = 24)											
Assemblies	91.7	100	95.8	87.5	95.8	100	100	91.7	95.8	100	100
PE reads	91.7	100	91.7	87.5	95.8	100	100	91.7	95.8	100	100
Total (n = 46)											
Assemblies	95.7	100	97.8	93.5	95.7	100	95.7	93.5	97.8	100	100
PE reads[e]	95.7	95.7	95.7	89.1	95.7	97.8	95.7	93.5	97.8	100	100

[a]Presence/absence of eight virulence genes from previously published WGS data (training set) or PCR (validation set).
[b]Multilocus sequence typing (MLST) results from previously published WGS data (training set) or Sanger sequencing (validation set).
[c]*rpoB* allelic typing (AT) results from previously published WGS data (training set) or Sanger sequencing (validation set).
[d]*panC* clade typing results from previously published WGS data.
[e]Illumina paired-end (PE) reads.

85.5%, respectively, when the default parameters for assembled genomes were used, and an overall sensitivity and specificity of 97.0% and 85.5%, respectively, when default parameters for Illumina paired-end reads were used.

4.4.2 Characteristics associated with *B. cereus* group phylogenetic clade III are most prevalent among genome assemblies currently available at NCBI

BTyper was used to perform virulence gene detection, MLST, *panC* clade typing, and *rpoB* allelic typing on 662 *B. cereus* group genome assemblies (157 assemblies labeled as *B. anthracis*, 353 assemblies as *B. cereus s.s.*, 2 assemblies as *B. cytotoxicus*, 19 assemblies as *B. mycoides*, 2 assemblies as *B. pseudomycoides*, 94 assemblies as *B. thuringiensis*, 3 assemblies as *B. toyonensis*, 21 assemblies as *B. weihenstephanensis*, and 11 assemblies as *B. wiedmannii*). Within the 662 assemblies, 13 virulence genes were detected in more than 90% of all genomes when the default minimum amino acid sequence identity and coverage thresholds of

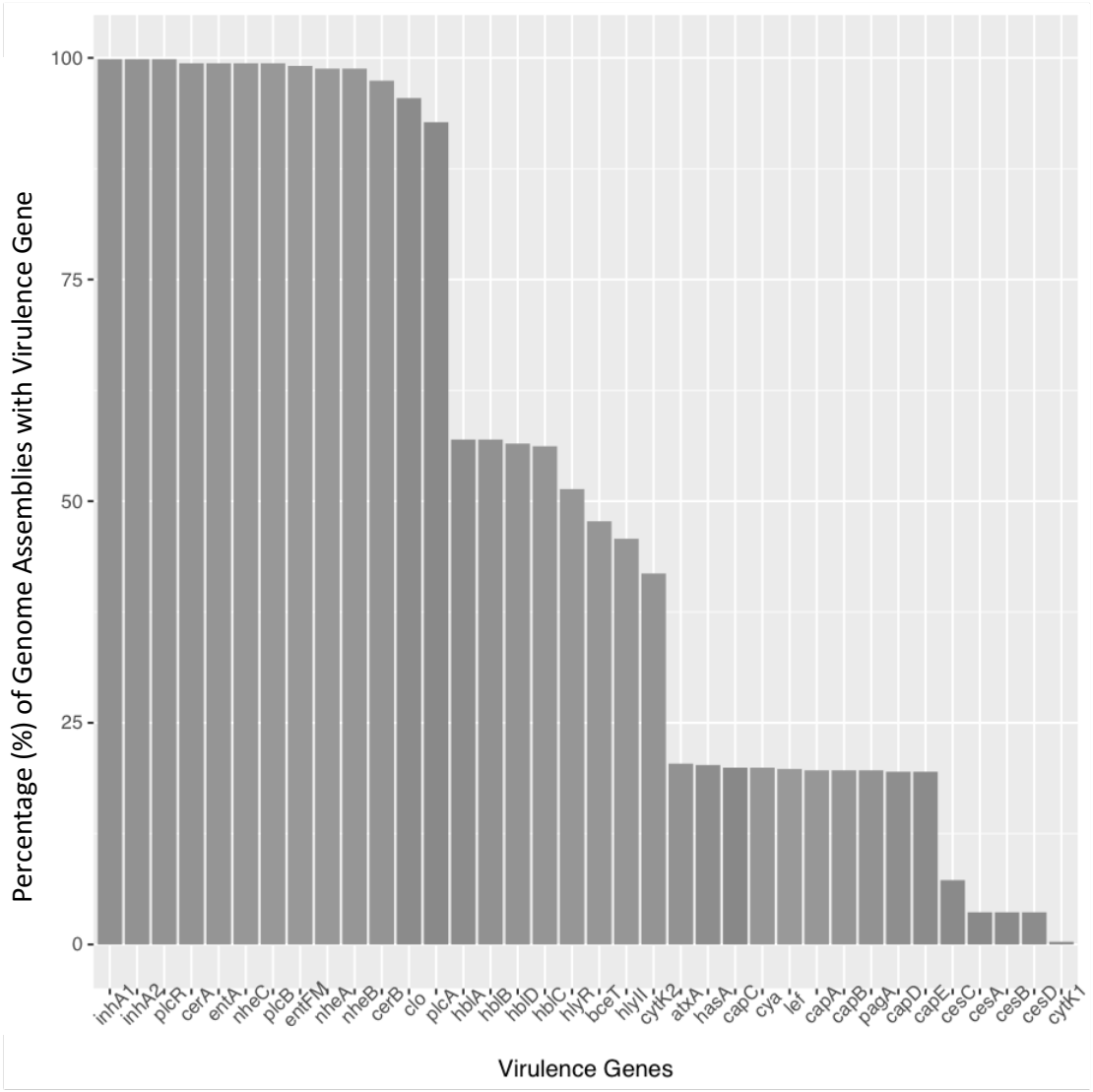


Figure 4.2: Percentage (%) of *B. cereus* group assemblies in which a particular virulence gene was detected. Minimum identity and coverage thresholds of 50 and 70%, respectively, were used for virulence gene detection.

50 and 70% were used, respectively (Figure 4.2). The least commonly detected gene was *cytK1* (Figure 4.2), which was detected in both available *B. cytotoxicus* genomes and no other WGS assemblies.

For *in silico* MLST, 544 assemblies were assigned to one of 213 *B. cereus* sequence types (STs), the most common of which was ST1 (n = 123 isolates). This was unsurprising, considering that ST1 is associated with *B. anthracis* (Helga-

son et al. 2004), and *B. anthracis* makes up a considerable portion (23.7%) of the *B. cereus* group genome assemblies currently in NCBI's database. *In silico rpoB* allelic typing grouped the 662 isolates into one of 43 different, best-matching *rpoB* allelic types (ATs), with 185 isolates matching AT463 most closely. AT463 has been previously associated with clade III isolates (Kovac et al. 2016), the phylogenetic clade that encompasses *B. anthracis*.

For *panC*-based phylogenetic clade typing, a *panC* locus was detected in 658 out of 662 genomes (Figure 4.3). The most commonly assigned clade was clade III, a polyphyletic clade which contains *B. anthracis*, as well as some strains currently misclassified in the NCBI database as *B. cereus s.s.* and *B. thuringiensis* (Kovac et al. 2016; Guinebretiere, Thompson, et al. 2008; Guinebretiere, Velge, et al. 2010). Together, clade IV, which consists of some *B. cereus s.s.* and *B. thuringiensis* strains (Kovac et al. 2016; Guinebretiere, Thompson, et al. 2008; Guinebretiere, Velge, et al. 2010), as well as the type strains of these two species, and clade III accounted for more than 75% of all *B. cereus* group WGS assemblies in the NCBI database (Figure 4.3). Clade VII, which contains the *B. cytotoxicus* (Guinebretiere, Auger, et al. 2013) type strain, was the most poorly represented clade; the two available *B. cytotoxicus* assemblies were placed here.

4.4.3 Application of BTyper to identify *B. anthracis*-associated genes in non-*anthracis Bacillus* isolates reveals virulence gene heterogeneity within genome assemblies from anthrax toxin-encoding isolates

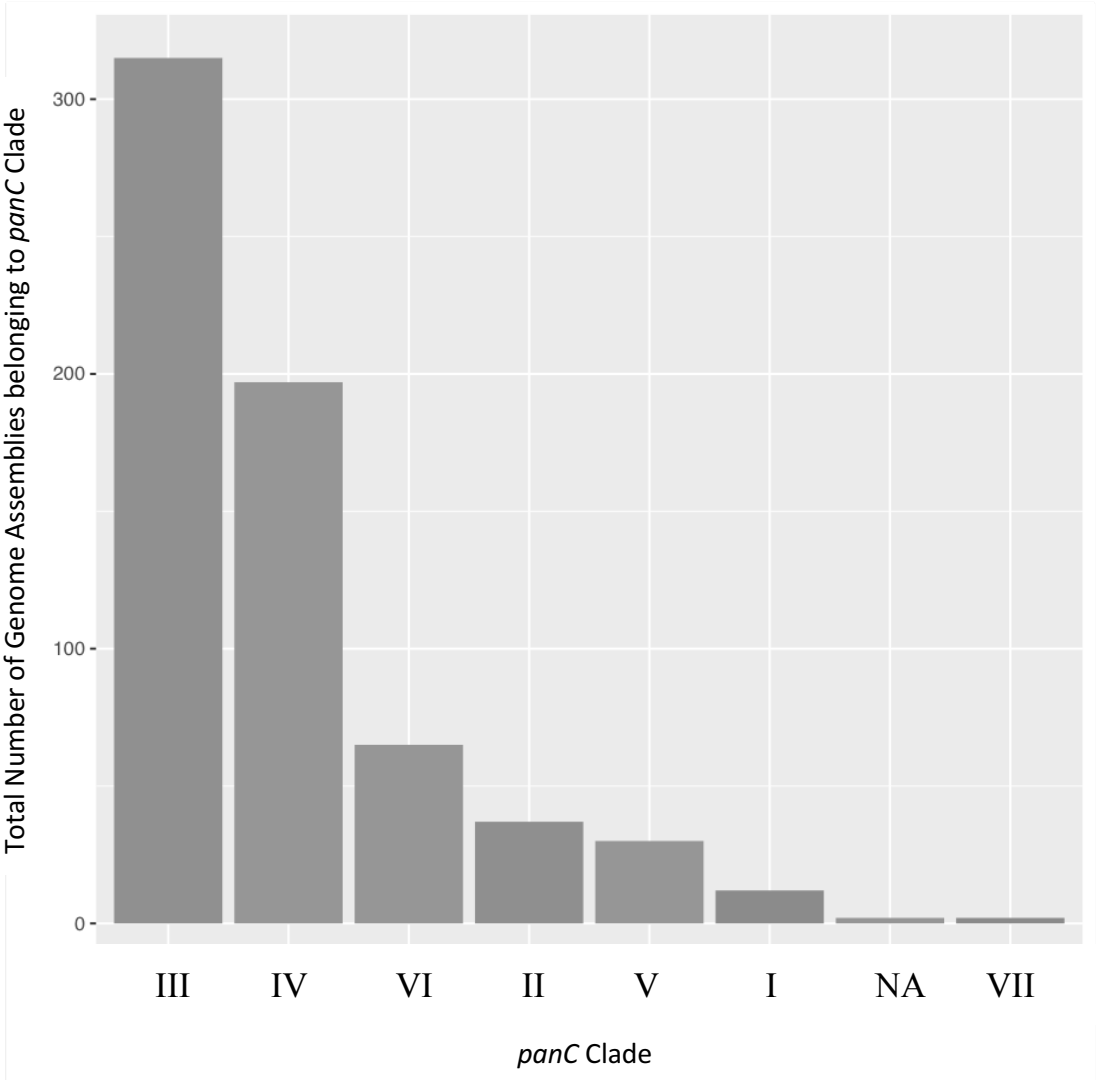


Figure 4.3: Closest-matching phylogenetic clade using the *panC* loci from 662 *B. cereus* group genome assemblies. A *panC* locus could not be assigned in 4 genome assemblies, which is denoted by NA.

When Fisher's exact test was used to determine if any virulence genes were significantly associated with a phylogenetic clade, virulence genes typically associated with *B. anthracis* were found to be significantly associated with members of clade III after a Bonferroni correction was applied ($P < 0.05$; Table 4.2). The *B. anthracis* toxin genes *cya* (edema factor-encoding), *lef* (lethal factor-encoding), and *pagA* (protective antigen-encoding), as well as their regulator gene *atxA*

(Dai et al. 1995), were found only in clade III isolates ($P < 0.05$; Table 4.2). In addition, *B. anthracis* polyglutamate capsule synthesis genes *capABCDE* (Candela, Mock, and Fouet 2005) were more commonly associated with clade III assemblies ($P < 0.05$; Table 4.2) and found primarily in genomes classified in the NCBI database as *B. anthracis*. Meanwhile, genes associated with diarrheal disease (Stenfors Arnesen, Fagerlund, and Granum 2008) were found to be significantly associated with clades II, IV, V, and VI ($P < 0.05$; Table 4.2); these included the diarrheal toxin genes *hblCDAB*, which were found to be significantly associated with clades II, IV, V, and VI ($P < 0.05$; Table 4.2), while being less common in members of clade III ($P < 0.05$; Table 4.2), driven by the large number of *B. anthracis* assemblies in this clade that did not possess these genes.

Table 4.2: Virulence genes significantly associated with 5 *B. cereus* group phylogenetic clades after a Bonferroni correction[a]

Clade	***Genes***
II	*hblCDAB*
III	*atxA*,[b] *capABCDE*, *cya*,[b] *hasA*, *hlyII*, *hlyR*, *lef*,[b] *pagA*[b]
IV	*bceT*, *cytK2*, *hblCDAB*
V	*bceT*, *hblCDAB*[c]
VI	*bceT*, *cesC*, *hblCDAB*[c]

[a]Significant at a P value of < 0.05. For exact corrected P values, see Table S7.
[b]Indicates a virulence gene that was detected only in its respective clade (includes clades I and VII).
[c]Indicates a virulence gene that was detected in all members of its respective clade.

Principal-component analysis (PCA) based on the presence/absence of virulence genes using BMiner revealed several assemblies labeled as *B. cereus* and *B. thuringiensis* that clustered with *B. anthracis* assemblies (Figure 4.4A). When k-medoids clustering was performed with an optimum k of 31, isolates classified in the NCBI database as *B. anthracis* were placed into clusters 1 through 8

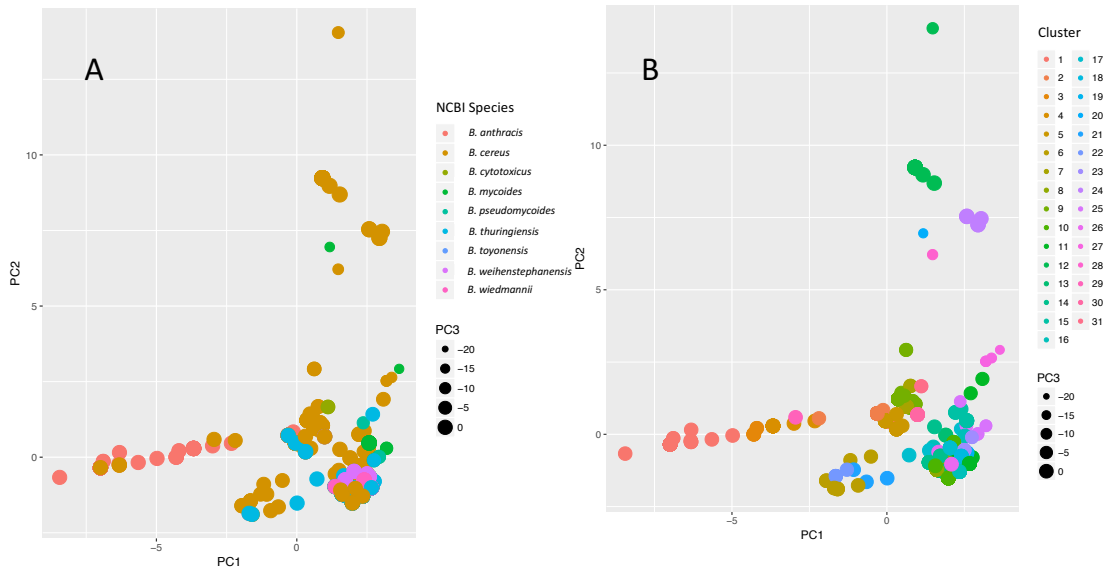


Figure 4.4: Principal-component analysis (PCA) of 662 *B. cereus* group genome assemblies based on presence/absence of virulence genes. Virulence gene typing was carried out using BTyper, while PCA was performed using BMiner. Principal components 1 (PC1) and 2 (PC2) are plotted on the x and y axes, respectively, while principal component 3 (PC3) corresponds to point size. Plots are colored by isolate species, as found in NCBI (A), and assigned cluster using *k*-medoids (B). To view interactive versions of these plots containing isolate names and metadata, all BTyper final results files and metadata can be downloaded from https://github.com/lmc297/BTyper/tree/master/sample_data and viewed in BMiner.

(Figure 4.4B). Additionally, clusters 17, 21, 22, and 29 did not contain any assemblies labeled in NCBI as *B. anthracis*, but they contained at least one assembly in which one or more of the *B. anthracis*-associated virulence genes identified using Fisher's exact test were detected (Figure 4.5).

Cluster 1 (Figure 4.4B), which contained the majority of isolates labeled as *B. anthracis*, contained 110 isolates, 107 of which were classified in the NCBI database as *B. anthracis*, and all of which belonged to *panC* clade III (Figure 4.5). Assemblies derived from human and veterinary clinical isolates associated with anthrax disease populated a large proportion of the cluster, including assemblies associated with isolates from the 2001 anthrax bioterrorism at-

			B. anthracis-Associated Genes										Emetic Toxin Genes				*cytK*		*bceT*	*hbl*				*hly*		*clo*	*plc*			*nhe*			*ent*		*cer*		*inhA*	
Cluster	Size	*panC*	*cya*	*lef*	*pagA*	*atxA*	*hasA*	*capA*	*capB*	*capC*	*capD*	*capE*	*cesA*	*cesB*	*cesC*	*cesD*	*cytK1*	*cytK2*	*bceT*	*hblA*	*hblB*	*hblC*	*hblD*	*hlyR*	*hlyII*	*clo*	*plcA*	*plcB*	*plcR*	*nheA*	*nheB*	*nheC*	*entA*	*entFM*	*cerA*	*cerB*	*inhA1*	*inhA2*
1	110	3	1.00	0.99	0.97	0.97	1.00	0.99	1.00	1.00	0.99	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.99	0.99	1.00	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.99
2	26	3, 4	0.00	0.00	0.00	0.00	0.00	0.04	0.04	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.00	1.00	1.00	0.96	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
3	6	3	0.00	0.00	0.00	1.00	0.00	1.00	1.00	1.00	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	0.67	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
4	18	3	0.94	0.94	1.00	0.94	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
5	26	3, 4*	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.27	0.00	0.00	0.00	0.00	1.00	0.96	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
6	10	3, 4	0.00	0.00	0.00	0.00	0.00	1.00	0.80	1.00	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.20	0.40	1.00	1.00	1.00	1.00	0.80	0.80	1.00	0.80	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
7	28	3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.93	1.00	1.00	1.00	1.00
8	40	2, 3, 4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	1.00	1.00	1.00	1.00	0.98	0.98	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
9	38	2, 3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.21	0.00	0.00	0.00	0.00	0.00	0.00	0.84	0.92	0.95	1.00	1.00	1.00	1.00	1.00	1.00	0.95	1.00	1.00	1.00
10	37	3, 4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.00	0.00	1.00	1.00	1.00	1.00	1.00	0.92	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
11	101	2, 3, 4, 5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	1.00	0.97	1.00	1.00	0.97	1.00	0.00	0.00	1.00	0.97	0.99	1.00	1.00	1.00	1.00	1.00	0.99	0.99	1.00	1.00	1.00
12	19	3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	0.00	0.11	0.05	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	0.95	0.95	0.95	1.00	1.00	1.00	1.00	1.00	1.00
13	20	2, 3, 4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.95	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
14	14	2, 3, 6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.07	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	0.79	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
15	14	2, 6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.79	1.00	1.00	1.00	1.00	0.00	0.00	0.93	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
16	25	4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.00	1.00	1.00	0.96	0.96	1.00	1.00	1.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.96	1.00	1.00	1.00	1.00
17	13	2, 3, 6	0.00	0.00	0.00	0.00	0.08	0.00	0.08	0.08	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	0.00	0.00	1.00	0.85	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.92	1.00	1.00
18	54	2, 4, 5, 6*	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
19	9	5, 6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
20	2	*	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	1.00	0.50	0.00	0.00	0.00	1.00	0.50	1.00	0.50	1.00	1.00
21	3	4, 5	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	0.67	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	0.67	1.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
22	5	3	1.00	1.00	1.00	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	1.00	1.00	1.00	1.00	1.00	0.80	1.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
23	9	1, 3, 6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	0.00	0.00	1.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.78	1.00	1.00
24	5	6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	0.00	0.00	0.60	1.00	1.00	1.00	1.00	0.00	0.00	1.00	0.80	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
25	7	1, 5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.71	1.00	1.00	0.86	0.86	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	0.57	1.00	1.00	0.00	1.00	1.00
26	7	4, 6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	0.86	1.00	1.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
27	5	1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	0.00	0.00	0.60	0.00	1.00	1.00	0.00	0.00	1.00	0.80	1.00	1.00	0.00	1.00	1.00
28	1	3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	1.00	1.00	1.00	0.00	1.00	0.00	0.00	0.00	0.00	1.00
29	1	3	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
30	7	2, 3, 4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
31	2	7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Figure 4.5: *k*-medoids clusters based on presence/absence of virulence genes detected using BTyper. Size corresponds to the number of assemblies assigned to a given cluster, while *panC* corresponds to *panC* clades found in the cluster, with an asterisk denoting one or more assemblies that could not be placed into a *panC* clade. Numbers within cells correspond to the proportion of assemblies in a given cluster in which the corresponding virulence gene was detected. Green shading corresponds to a virulence gene detected in more than 90% of all assemblies in a cluster, while red shading corresponds to a virulence gene detected in fewer than 10% of all assemblies in a cluster. Yellow shading corresponds to *B. anthracis*-associated genes detected in fewer than 90% but greater than 0% of assemblies in a cluster.

tacks (https://www.ncbi.nlm.nih.gov/bioproject/299), European heroin users and an associated outbreak (Ruckert et al. 2012; Price et al. 2012), and a 2011 outbreak in Swedish cattle (Agren et al. 2014). Three assemblies labeled as *B. cereus* clustered among them (Figure 4.4B). Two of these assemblies were labeled as *B. cereus* strain 03BB102, an isolate that was thought to cause fatal pneumonia in a welder in San Antonio, TX (Table 4.3), while the third was labeled as *B. cereus* biovar anthracis strain CI, which caused fatal anthrax in a chimpanzee (Table 4.3) (Klee et al. 2010). Consistent with these findings, placement into cluster 1 was driven largely by an assembly's possession of all, or nearly all, anthrax-associated genes identified using Fisher's exact test (Figure 4.6); the anthrax toxin genes *cya*, *lef*, and *pagA*, toxin regulator gene *atxA*, hyaluronic acid capsule gene *hasA*, and *B. anthracis* polyglutamate capsule genes *capABCDE* were detected in nearly all (> 97%) cluster 1 assemblies (Figure 4.5).

Despite the fact that all assemblies classified in NCBI as *B. anthracis* were assigned to clusters 1 through 8, the only other clusters in addition to cluster 1 in which anthrax toxin genes were detected were clusters 4 and 22. Like cluster 1, all isolates in clusters 4 and 22 belonged to *panC* clade III, and nearly all possessed the anthrax toxin genes *cya*, *lef*, and *pagA*, regulator gene *atxA*, and hyaluronic acid capsule gene *hasA* (Figure 4.5). However, the *B. anthracis* polyglutamate capsule genes *capABCDE* were not detected in any of the cluster 4 or cluster 22 assemblies at the default identity and coverage thresholds (Figure 4.5). While cluster 4 (n = 18; Figure 4.4B) contained only isolates classified in the NCBI database as *B. anthracis*, it contained assemblies from several strains with attenuated virulence, including several vaccine strains (Lekota et al. 2015; Okinaka et al. 2014) (https://www.ncbi.nlm.nih.gov/biosample/SAMN06270273/). Cluster 22 (n =

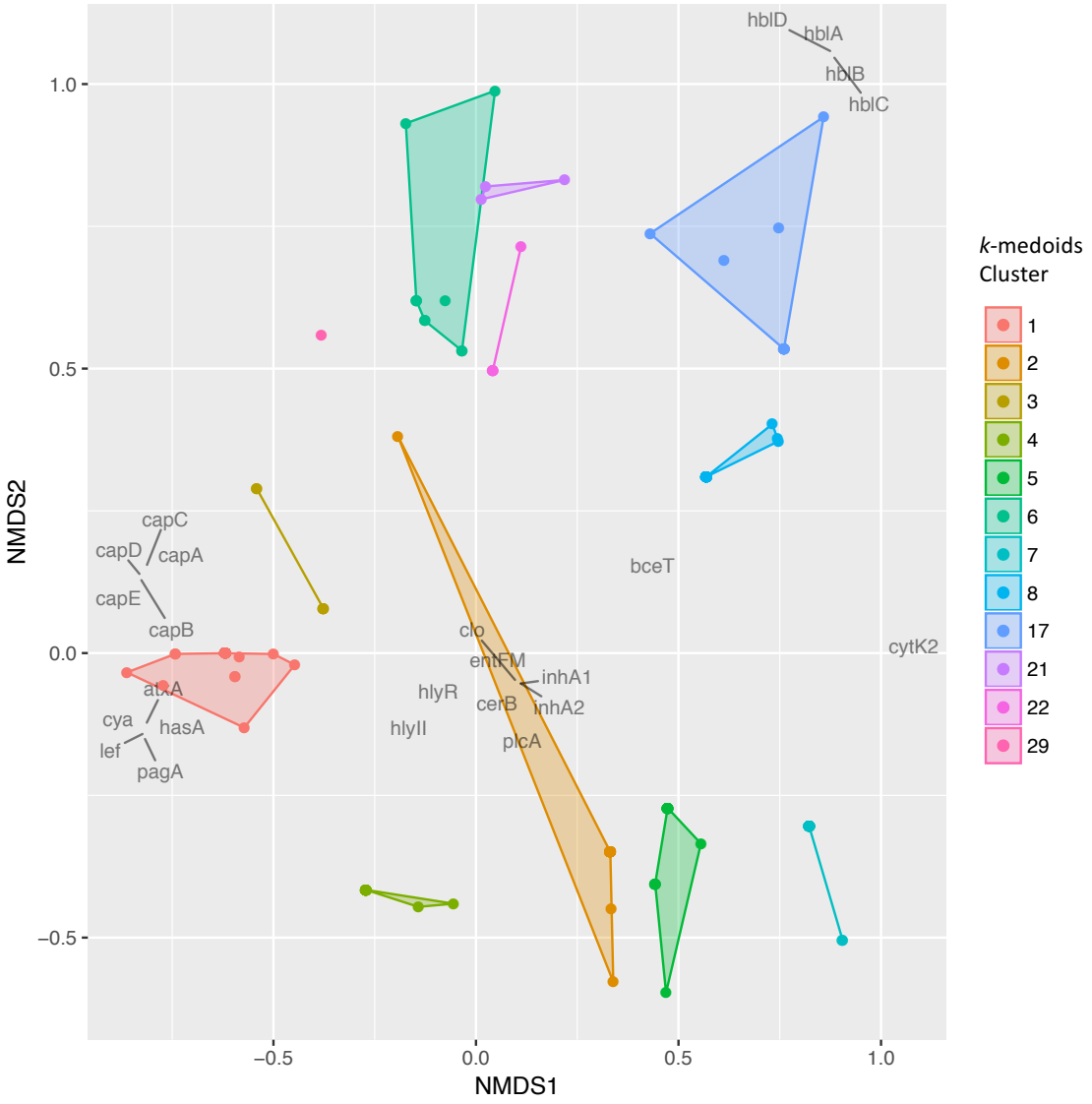


Figure 4.6: Nonmetric multidimensional scaling (NMDS) plot of *Bacillus cereus* group clusters that (i) possessed at least one assembly that was classified as *Bacillus anthracis* in NCBI, and/or (ii) possessed at least one assembly in which at least one *B. anthracis*-associated virulence gene (*cya*, *lef*, *pagA*, *atxA*, *hasA*, and/or *capABCDE*) was detected using BTyper. NMDS was performed in BMiner using virulence gene presence/absence data and a Jaccard dissimilarity metric. Isolates are represented by points, and convex hulls and shading correspond to the assigned *k*-medoids cluster. Virulence genes are plotted in dark gray.

Table 4.3: Non-*anthracis Bacillus* assemblies in which anthrax toxin genes *cya*, *lef*, and/or *pagA* were detected using BTyper

Cluster[a]	*NCBI species classification*	*panC clade*[b]	*GenBank accession no.*[c]	*Strain*	*Isolate source (reference)*	Gene(s) detected?					
						cya	*lef*	*pagA*	*atxA*	*hasA*	*capABCDE*
1	*B. cereus*	III	GCA_000022505.1, GCA_000832405.1	03BB102	Human with fatal pneumonia, San Antonio, TX, USA[d]	+	+	+	-	+	+
1	*B. cereus*	III	GCA_000143605.1	Biovar anthracis strain CI	Chimpanzee with fatal anthrax, Ivory Coast[e]	+	+	+	+	+	+
22	*B. cereus*	III	GCA_000167215.1, GCA_000832805.1	G9241	Human with pneumonia, nausea, and vomiting, LA, USA[f]	+	+	+	+	+	-
22	*B. cereus*	III	GCA_000688755.1	BcFL2013	Human with anthrax-like skin lesion, FL, USA[g]	+	+	+	+	+	-
22	*B. cereus*	III	GCA_000789315.1	03BB87	Human with fatal pneumonia, Lubbock, TX, USA[h]	+	+	+	+	+	-
22	*B. cereus*	III	GCA_002007005.1	LA2007	Human with fatal pneumonia and septic shock, Galliano, LA, USA[i]	+	+	+	+	+	-

[a]Clusters were assigned using a *k*-medoids approach ($k = 31$).
[b]*panC* clades were assigned using BTyper.
[c]Multiple accession numbers are given for strains associated with multiple assemblies.
[d]https://www.ncbi.nlm.nih.gov/bioproject/31307
[e] (Klee et al. 2010)
[f] (Alex R. Hoffmaster et al. 2004)
[g] (Gee et al. 2014)
[h] (Johnson et al. 2015)
[i] (Pena-Gonzalez et al. 2017)

5; Figure 4.4B), however, contained 5 anthrax-associated assemblies, all of which were classified in the NCBI database as *B. cereus* (Table 4.3). All assemblies in cluster 22 originated from human clinical isolates in which the isolate was classified as *B. cereus*, but the patient presented anthrax-like symptoms; two assemblies were of *B. cereus* strain G9241, a strain of *Bacillus* isolated from the sputum and blood of a patient with pneumonia, nausea, and vomiting (Alex R. Hoffmaster et al. 2004). The isolate, which had been classified as *B. cereus* via biochemical tests and 16S rRNA gene sequencing, was found to possess the anthrax toxin gene *pagA* but not the polyglutamate capsule genes *capABCDE* (Alex R. Hoffmaster et al. 2004), which is consistent with its classification using BTyper (Table 4.3). BTyper's classification of the three other assemblies in this cluster also aligned with their previously published descriptions and included

the following: (i) a *B. cereus* assembly associated with an isolate from a patient in Florida possessing an anthrax-like skin lesion (Gee et al. 2014), which was found to possess anthrax toxin genes *cya*, *lef*, and *pagA* and the hyaluronic acid capsule gene *hasA* and belong to ST78 (Gee et al. 2014), (ii) a *B. cereus* isolate from a patient with a fatal case of pneumonia in Lubbock, TX, that was also found to possess *B. anthracis* virulence genes (Johnson et al. 2015), and (iii) an assembly associated with a *B. cereus* isolate that was found to possess anthrax toxin genes and *hasA* and was isolated from a patient in Galliano, LA, who had a fatal case of pneumonia and septic shock (Table 4.3) (Pena-Gonzalez et al. 2017).

While no anthrax toxin genes were detected outside clusters 1, 4, and 22, other *B. anthracis*-associated genes identified using Fisher's exact test were detected in several other clusters and assemblies. Cluster 3 ($n = 6$; Figure 4.4B) contained 6 *B. anthracis* assemblies belonging to *panC* clade III in which the *B. anthracis* toxin regulator gene *atxA* and polyglutamate capsule genes *capABCDE* were detected (Figure 4.5). Other assemblies in this cluster included *B. anthracis* strain Smith 1013, described as "Pasteur-like" in that it possessed plasmid pXO2 (the plasmid associated with *cap* genes) but not plasmid pXO1 (the plasmid associated with *B. anthracis* toxin genes) (Rasko et al. 2005; Terzi et al. 2014), as well as *B. anthracis* strain Pasteur itself (Table 4.4).

The polyglutamate capsule genes *capABCDE* were also detected in assemblies assigned to clusters 6, 21, and 29 (Table 4.4). Cluster 6 ($n = 10$; Figure 4.4B) contained 10 assemblies: 1 assembly classified in NCBI as *B. anthracis*, 7 assemblies classified as *B. cereus*, and 2 assemblies classified as *B. thuringiensis*. Members of this cluster belonged to *panC* clades III and IV, and consistent with

Table 4.4: Non-*anthracis Bacillus* assemblies in which *B. anthracis*-associated genes were detected, excluding anthrax toxin genes *cya*, *lef*, and *pagA* and regulator *atxA*

						Gene(s) detected?					
Cluster	*NCBI species classification*	*panC clade*	*GenBank accession no.*[a]	*Strain*	*Isolate source (reference)*	*hasA*	*capA*	*capB*	*capC*	*capD*	*capE*
2	*B. cereus*	III	GCA_001286905.1	JRS1	Rhazya stricta rhizosphere, Jeddah, Saudi Arabia[b]	-	+	+	+	-	-
6	*B. cereus*	III	GCA_000003955.1	AH1273	Human blood, Iceland[c]	-	+	+	+	+	+
6	*B. cereus*	III	GCA_000161395.1	AH1272	Amniotic fluid, Iceland[c]	-	+	-	+	+	+
6	*B. cereus*	III	GCA_000181655.1, GCA_000832865.1	03BB108	Dust containing pneumonia-causing *B. cereus* strain 03BB012[d]	-	+	+	+	+	+
6	*B. cereus*	IV	GCA_000398945.1	Schrouff	Food[e]	-	+	+	+	+	+
6	*B. cereus*	IV	GCA_000399185.1	K-5975c	Food[e]	-	+	+	+	+	+
6	*B. cereus*	IV	GCA_000399305.1	HuB4-4	Soil, Belgium[e]	-	+	-	+	+	+
6	*B. thuringiensis*	III	GCA_000161595.1	Serovar Monterrey strain BGSC 4AJ1	Mexico[f]	-	+	+	+	+	+
6	*B. thuringiensis*	IV	GCA_001640965.1	BGSC 4C1	*Bombyx mori*, Czechoslovakia[g]	-	+	+	+	+	+
17	*B. cereus*	VI	GCA_002014585.1	FSL H8-0485	Soil, USA[h]	+	-	-	-	-	-
17	*B. thuringiensis*	III	GCA_000948155.1	Et10/1	Geothermal spring, Lirima thermal springs, Chile[i]	-	-	+	+	-	-
21	*B. cereus*	IV	GCA_000161315.1	F65185	Open fracture, NY, USA[j]	-	+	+	+	+	+
21	*B. cereus*	V	GCA_000290835.1	VD115	Soil, Guadeloupe[e]	-	+	+	+	+	+
21	*B. thuringiensis*	IV	GCA_001677055.1	BGSC 4BT1	Red soil, China[k]	-	+	+	+	+	-
29	*B. cereus*	III	GCA_001913295.1	MOD1_Bc119	Whole black pepper, USA[l]	-	+	+	+	+	+

[a]Multiple accession numbers are given for strains associated with multiple assemblies.
[b]https://www.ncbi.nlm.nih.gov/bioproject/290051
[c] (Zwick et al. 2012)
[d]https://www.ncbi.nlm.nih.gov/bioproject/19959
[e] (Van der Auwera et al. 2013)
[f]https://www.ncbi.nlm.nih.gov/bioproject/29709
[g]https://www.ncbi.nlm.nih.gov/biosample/SAMN04628222/
[h]https://www.ncbi.nlm.nih.gov/biosample/SAMN06242081
[i]https://www.ncbi.nlm.nih.gov/biosample/SAMN03025783
[j]https://www.ncbi.nlm.nih.gov/bioproject/29689
[k]https://www.ncbi.nlm.nih.gov/biosample/SAMN04000100; *capE* was detected at a lower amino acid identity (47.7%, compared to the default threshold of 50%)
[l]https://www.ncbi.nlm.nih.gov/biosample/SAMN05608051

the detection of *cap* genes in this cluster, one of the *B. thuringiensis* assemblies in this group had been shown to produce a polyglutamate capsule (Cachat et al. 2008). Cluster 21 (n = 3; Figure 4.4B) contained 2 assemblies labeled as *B. cereus* and 1 assembly labeled as *B. thuringiensis*. One of the *B. cereus* assemblies came from *B. cereus* strain F65185, which was confirmed to belong to ST168 and was isolated from a patient in New York with an open fracture wound (Table 4.4). Members of this group belonged to either *panC* clade IV or V. Cluster 29 (n = 1;

Figure 4.4B) consisted of a single *B. cereus* assembly belonging to *panC* clade III and associated with a strain isolated from whole black pepper in the United States in 2015 (Table 4.4).

Additionally, *cap* genes were detected in a single isolate in clusters 2 and 17 (n = 26 and 13, respectively; Figure 4.4B). However, *B. anthracis*-associated genes were not detected in any other assemblies in this cluster, despite being composed primarily of assemblies classified as *B. anthracis* (21, 4, and 1 assemblies labeled in NCBI as *B. anthracis*, *B. cereus*, and *B. thuringiensis*, respectively). Consistent with a lack of virulence genes, this cluster contained the genome of the avirulent strain *B. anthracis* Ames, which is commonly used in laboratory settings and does not possess *B. anthracis* plasmid pXO1 or pXO2 (https://www.ncbi.nlm.nih.gov/bioproject/57909). All non-*anthracis* *Bacillus* assemblies in this group were isolated from either food or environmental sources, and all belonged to either *panC* clade III or IV.

4.4.4 Application of BTyper to identify assemblies associated with emetic *B. cereus* group isolates

Assemblies possessing emetic toxin genes *cesABCD* were grouped into two clusters using *k*-medoids. Cluster 12 (n = 19; Figure 4.4B) consisted of 19 assemblies classified as *B. cereus* in NCBI. All belonged to *panC* clade III, *cesABCD* were detected in all assemblies, and *hblCDAB* were not detected in any assemblies (Figure 4.5). Included in this cluster was strain AH187, an isolate from the United Kingdom that was responsible for a 1972 emetic outbreak (Table 4.5). This isolate tested positive for emetic toxin (cereulide) for-

mation and nonhemolytic enterotoxin (Nhe) and negative for Hbl hemolytic enterotoxin and cytotoxin K, and it belonged to MLST ST26 (Table 4.5) (https://www.ncbi.nlm.nih.gov/bioproject/17715); these findings were confirmed using BTyper. Other notable strains in this cluster included (i) emetic strain *B. cereus* H3081.97, a *B. cereus* strain of sequence type 144 (ST144) which is closely related to strain AH187, and (ii) emetic strain *B. cereus* NC7401 (Takeno et al. 2012).

Table 4.5: *B. cereus* group assemblies in which emetic toxin genes *cesABCD* were detected.

Cluster	*NCBI species classification*	*panC clade*	*GenBank accession no.*	*Strain*	*Isolate source (reference)*
12	*B. cereus*	III	GCA_000021225.1	AH187	Vomit of a person who ate cooked rice; isolate was associated with an emetic outbreak in 1972 (https://www.ncbi.nlm.nih.gov/bioproject/17715)
12	*B. cereus*	III	GCA_000161075.1	BDRD-ST26	BDRD stock strain (Zwick et al. 2012) [a]
12	*B. cereus*	III	GCA_000171035.2	H3081.97	Food; emetic toxin-producing isolate from 1997 outbreak linked to rice, TX, USA
12	*B. cereus*	III	GCA_000283675.1	NC7401	Emetic isolate (Takeno et al. 2012)
12	*B. cereus*	III	GCA_000290935.2	IS075	Wild mammal (vole) (Ladeuze et al. 2011)
12	*B. cereus*	III	GCA_000290995.1	AND1407	Black currant (Hoton et al. 2009) (53)
12	*B. cereus*	III	GCA_000291235.1	MSX-A12	Not available (Van der Auwera et al. 2013)
12	*B. cereus*	III	GCA_000399205.1	IS845/00	Bank vole, Poland (Van der Auwera et al. 2013; I. Swiecicka and De Vos 2003)
12	*B. cereus*	III	GCA_000399225.1	IS195	Bank vole, Poland (Van der Auwera et al. 2013; I. Swiecicka and De Vos 2003)
12	*B. cereus*	III	GCA_000743195.1	F1-15	Foodborne source (Zhong et al. 2007)
12	*B. cereus*	III	GCA_001566375.1	MB.15	Food, Munich, Germany (Crovadore et al. 2016)
12	*B. cereus*	III	GCA_001566385.1	MB.18	Food, Munich, Germany (Crovadore et al. 2016)
12	*B. cereus*	III	GCA_001566435.1	MB.16	Food, Munich, Germany (Crovadore et al. 2016)
12	*B. cereus*	III	GCA_001566445.1	MB.17	Food, Munich, Germany (Crovadore et al. 2016)
12	*B. cereus*	III	GCA_001566455.1	MB.21	Food, Munich, Germany (Crovadore et al. 2016)
12	*B. cereus*	III	GCA_001566465.1	MB.8	Food, Munich, Germany (Crovadore et al. 2016)
12	*B. cereus*	III	GCA_001566515.1	MB.8-1	Food, Munich, Germany (Crovadore et al. 2016)
12	*B. cereus*	III	GCA_001566525.1	MB.20	Food, Munich, Germany (Crovadore et al. 2016)
12	*B. cereus*	III	GCA_001566535.1	MB.22	Food, Munich, Germany (Crovadore et al. 2016)
24	*B. cereus*	VI	GCA_000291155.1	MC67	Sandy loam, Denmark (Thorsen et al. 2006; Van der Auwera et al. 2013; Hendriksen, Hansen, and Johansen 2006)
24	*B. cereus*	VI	GCA_000291315.1	CER074	Raw milk (Hoton et al. 2009)
24	*B. cereus*	VI	GCA_000291335.1	CER057	Parsley (Hoton et al. 2009)
24	*B. cereus*	VI	GCA_000293605.1	BtB2-4	Forest soil (Hoton et al. 2009)
24	*B. cereus*	VI	GCA_000399245.1	MC118	Sandy loam, Denmark (Thorsen et al. 2006; Van der Auwera et al. 2013; Hendriksen, Hansen, and Johansen 2006)

[a] BDRD, Biological Defense Research Directorate

The other cluster in which all *cesABCD* genes were detected in all assemblies was cluster 24 (n = 5; Figure 4.4B). This cluster contained 5 assemblies classified as *B. cereus*, all of which belonged to *panC* clade VI (Table 4.5). Unlike cluster 12, *hblCDAB* genes were detected in all assemblies in this cluster (Figure 4.5). The assemblies in this cluster originated from food and environmental isolates (Table 4.5). Despite their assemblies being classified in the NCBI database as *B. cereus*, all 5 strains in this cluster were classified as emetic *B. weihenstephanensis* in their respective manuscripts, and all were capable of growth at 8°C (Hoton et al. 2009; Thorsen et al. 2006).

4.5 Discussion

4.5.1 Accessible whole-genome sequence analysis tools can facilitate improved taxonomic classification and characterization of *B. cereus* group isolate virulence potential

As whole-genome sequencing becomes more widely used in the realms of public health and food safety, the ability to classify potential pathogenic microorganisms quickly and effectively becomes increasingly important. A number of bioinformatics tools already exist for this purpose, including SRST2, which can be used to perform MLST and detect antimicrobial resistance genes using Illumina reads (Inouye et al. 2014); SeqSero, which performs *in silico* serotyping using Illumina reads or nucleotide assemblies from *Salmonella enterica* isolates (Zhang et al. 2015); PlasmidFinder, which can be used to detect plasmids in iso-

lates using Illumina reads or nucleotide assemblies (Carattoli et al. 2014); and VirulenceFinder, which can be used to detect virulence genes in *Listeria monocytogenes*, *Staphylococcus aureus*, *Escherichia coli*, and *Enterococcus* (Joensen et al. 2014). Recently, methods such as *in silico* MLST and virulence gene detection have been combined into single computational pipelines that can be used to characterize numerous bacterial species (Thomsen et al. 2016). Here, we have created a bioinformatics tool specific to the *Bacillus cereus* group that combines virulence gene detection using a curated database of *B. cereus* virulence factors with *in silico* manifestations of established molecular and virulence typing methods to phylogenetically classify and rapidly assess the virulence potential of any *B. cereus* group isolate. Additionally, we have provided a companion application, BMiner, that allows users to interact with data from hundreds of genomes at once, which we anticipate will become increasingly valuable as more *B. cereus* group genomes are sequenced.

The *in silico* typing methods employed by BTyper and other bioinformatics tools are valuable from a public health and food safety perspective, due to their (i) speed, as BTyper and similar tools can be used to perform gene detection and typing tasks in seconds using assembled genomes (Zhang et al. 2015; Carattoli et al. 2014); (ii) scalability, with the ability to provide users with information about a single isolate or hundreds from the command line (Inouye et al. 2014; Zhang et al. 2015); and (iii) ability to output concise and easily interpretable summaries of large amounts of data (Inouye et al. 2014), making it easy for a user to understand their results, share data with colleagues, and make informed decisions about an isolate in question (i.e., is it pathogenic or not). Additionally, the use of virulence gene-based typing as employed by BTyper offers the advantage that isolates can be classified according to their virulence poten-

tial, which means that one does not have to make any prior assumptions about the taxonomic classification of an isolate in question. This marks a valuable step forward in distinguishing pathogenic *B. cereus* group isolates from their nonpathogenic counterparts; however, marked improvements could be made to BTyper and similar tools through the integration of phenotypic data. By associating genotypic characteristics of *B. cereus* group isolates with phenotypic data, such as host illness and symptoms and growth temperature, BTyper and other tools used to genotype foodborne pathogens may become more valuable from a risk assessment perspective.

4.5.2 Analysis of publicly available *B. cereus* group assemblies using BTyper and BMiner identifies virulence gene-based clusters that capture phylogenetic heterogeneity in isolates with similar phenotypes

Using the output of BTyper and BMiner, virulence gene profiles of 662 *B. cereus* group genomes were assigned to one of 31 clusters by employing a *k*-medoids approach, without making unnecessary prior assumptions about an assembly's taxonomic classification in the public domain. This allowed for the identification of several well-defined clusters with clinical or taxonomic relevance, including (i) fully virulent *B. anthracis* and *B. anthracis*-like *B. cereus* (cluster 1), (ii) *capABCDE*-negative anthrax-causing *B. cereus* strains (cluster 22), (iii) *B. anthracis* with attenuated virulence (clusters 3 and 4), (iv) 2 emetic clusters (clusters 12 and 24), and (v) *B. cytotoxicus* (cluster 31). The clustering of the emetic

assemblies into 2 separate clusters reflected the observed heterogeneity among emetic strains of *B. cereus* and *B. weihenstephanensis*: Hoton et al. (Hoton et al. 2009) described two distinct clusters formed by emetic toxin-producing *B. cereus* group strains, with psychrotolerant *B. weihenstephanensis* strains belonging to a distinct emetic cluster (referred to in its respective manuscript as cluster II) (Hoton et al. 2009; Castiaux et al. 2014). Assemblies from these strains were placed into a single cluster (*k*-medoids cluster 24) consisting of *B. weihenstephanensis* assemblies belonging to *panC* clade VI, while members of Hoton et al.'s emetic cluster I were placed into a second cluster (*k*-medoids cluster 12) containing assemblies belonging to *panC* clade III. For *B. cytotoxicus*, the two available assemblies, both of which were the only *panC* clade VII representatives, were placed into a single cluster composed of only themselves (*k*-medoids cluster 31), driven largely by their possession of *cytK1*, as described by Guinebretire et al. (Guinebretiere, Velge, et al. 2010). For *B. anthracis*, strains possessing both anthrax virulence plasmids (pXO1 and pXO2) were assigned to cluster 1, distinguishing them from attenuated strains in which one or neither plasmid was detected, as well as *B. cereus* strains that caused anthrax-like disease (cluster 22). Despite lacking the polyglutamate capsule genes *capABCDE*, *B. cereus* strains in cluster 22 were able to cause anthrax-like symptoms using a second capsule encoded by *B. cereus* exopolysaccharide genes *bpsXABCDEFGH* (*bpsX-H*) on a different plasmid, pBC218 (Oh et al. 2011). The *bpsX-H* operon in its entirety was detected in 4 of the 5 anthrax-causing, *capABCDE*-negative *B. cereus* assemblies in cluster 22 (all but strain BcFL2013) and in no other cluster. It is likely that results like this from additional studies will be able to further resolve clade assignments and disease phenotypes with BTyper; recently, Bazinet identified numerous genes associated with phenotypic traits, such as anthrax and food

poisoning (Bazinet 2017). Here, we found associations between *B. cereus* group virulence genes and the *panC* clade, and virulence gene heterogeneity within disease phenotypes was identified. As more *B. cereus* group WGS and associated metadata become available, the potential for identifying new virulence alleles or phylogenetic markers that can further identify alleles or genes that are not only associated with a particular disease, but with specific symptoms or a clinical outcome using BTyper, becomes promising. For example, future work will be needed to better define specific genetic markers that can classify *B. cereus* group strains and clusters that are likely to cause diarrheal illnesses. Future epidemiological studies that assess the associations between different clusters and disease outcomes and symptoms will also provide an opportunity to further define and refine the types of disease outcomes and public health risks associated with different *B. cereus* group strains.

4.6 Acknowledgments

This material is based on work supported by the National Science Foundation Graduate Research Fellowship Program under grant no. DGE-1144153. Partial funding for this project was provided by the New York State Dairy Promotion Advisory Board through the New York State Department of Agriculture and Markets.

4.7 References

Aceves-Diez, Angel E., Kelly J. Estrada-Castaneda, and Laura M. Castaneda-Sandoval (2015). "Use of *Bacillus thuringiensis* supernatant from a fermen-

tation process to improve bioremediation of chlorpyrifos in contaminated soils". In: *Journal of Environmental Management* 157, pp. 213–219.

Agren, Joakim, Maria Finn, Bjorn Bengtsson, and Bo Segerman (2014). "Microevolution during an Anthrax Outbreak Leading to Clonal Heterogeneity and Penicillin Resistance". In: *PLOS ONE* 9.2, pp. 1–7. DOI: 10.1371/journal.pone.0089112.

Ammons, David R. et al. (2016). "Anti-cancer Parasporin Toxins are Associated with Different Environments: Discovery of Two Novel Parasporin 5-like Genes". In: *Current Microbiology* 72, pp. 184–189. DOI: 10.1007/s00284-015-0934-3.

Armada, Elisabeth, Rosario Azcon, Olga M. Lopez-Castillo, Monica Calvo-Polanco, and Juan Manuel Ruiz-Lozano (2015). "Autochthonous arbuscular mycorrhizal fungi and *Bacillus thuringiensis* from a degraded Mediterranean area can be used to improve physiological traits and performance of a plant of agronomic interest under drought conditions". In: *Plant Physiology and Biochemistry* 90, pp. 64–74.

Bache, Stefan Milton and Hadley Wickham (2014). *magrittr: A Forward-Pipe Operator for R*. R package version 1.5.

Bankevich, A. et al. (2012). "SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing". In: *J Comput Biol* 19.5, pp. 455–77. DOI: 10.1089/cmb.2012.0021.

Barrett, T. et al. (2012). "BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata". In: *Nucleic Acids Res* 40.Database issue, pp. D57–63. DOI: 10.1093/nar/gkr1163.

Bazinet, Adam L. (2017). "Pan-genome and phylogeny of *Bacillus cereus sensu lato*". In: *BMC evolutionary biology* 17.1, pp. 176–176. DOI: 10.1186/s12862-017-1020-1.

Bohm, M. E., C. Huptas, V. M. Krey, and S. Scherer (2015). "Massive horizontal gene transfer, strictly vertical inheritance and ancient duplications differen-

tially shape the evolution of *Bacillus cereus* enterotoxin operons *hbl*, *cytK* and *nhe*". In: *BMC Evol Biol* 15, p. 246. DOI: 10.1186/s12862-015-0529-4.

Caamano-Antelo, S. et al. (2015). "Genetic discrimination of foodborne pathogenic and spoilage *Bacillus* spp. based on three housekeeping genes". In: *Food Microbiology* 46, pp. 288–298.

Cachat, Elise, Margaret Barker, Timothy D. Read, and Fergus G. Priest (2008). "A *Bacillus thuringiensis* strain producing a polyglutamate capsule resembling that of *Bacillus anthracis*". In: *FEMS Microbiology Letters* 285.2, pp. 220–226. DOI: 10.1111/j.1574-6968.2008.01231.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1574-6968.2008.01231.x.

Camacho, C. et al. (2009). "BLAST+: architecture and applications". In: *BMC Bioinformatics* 10, p. 421. DOI: 10.1186/1471-2105-10-421.

Candela, Thomas, Michele Mock, and Agnes Fouet (2005). "CapE, a 47-amino-acid peptide, is necessary for *Bacillus anthracis* polyglutamate capsule synthesis". In: *Journal of bacteriology* 187.22, pp. 7765–7772. DOI: 10.1128/JB.187.22.7765-7772.2005.

Carattoli, A. et al. (2014). "*In silico* detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing". In: *Antimicrob Agents Chemother* 58.7, pp. 3895–903. DOI: 10.1128/AAC.02412-14.

Cardazzo, B. et al. (2008). "Multiple-locus sequence typing and analysis of toxin genes in *Bacillus cereus* food-borne isolates". In: *Appl Environ Microbiol* 74.3, pp. 850–60. DOI: 10.1128/AEM.01495-07.

Castiaux, V. et al. (2014). "Diversity of pulsed-field gel electrophoresis patterns of cereulide-producing isolates of *Bacillus cereus* and *Bacillus weihenstephanensis*". In: *FEMS Microbiol Lett* 353.2, pp. 124–31. DOI: 10.1111/1574-6968.12423.

CDC. *Anthrax*. https://www.cdc.gov/anthrax/index.html.

Chang, Winston, Joe Cheng, JJ Allaire, Yihui Xie, and Jonathan McPherson (2017). *shiny: Web Application Framework for R*. R package version 1.0.1.

Chen, M.L. and H.Y. Tsen (2002). "Discrimination of *Bacillus cereus* and *Bacillus thuringiensis* with 16S rRNA and *gyrB* gene based PCR primers and sequencing of their annealing sites". In: *Journal of Applied Microbiology* 92.5, pp. 912–919. DOI: 10.1046/j.1365-2672.2002.01606.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1046/j.1365-2672.2002.01606.x.

Cock, P. J. et al. (2009). "Biopython: freely available Python tools for computational molecular biology and bioinformatics". In: *Bioinformatics* 25.11, pp. 1422–3. DOI: 10.1093/bioinformatics/btp163.

Cole, James R. et al. (2014). "Ribosomal Database Project: data and tools for high throughput rRNA analysis". In: *Nucleic acids research* 42.Database issue, pp. D633–D642. DOI: 10.1093/nar/gkt1244.

Crovadore, Julien et al. (2016). "Whole-Genome Sequences of Seven Strains of *Bacillus cereus* Isolated from Foodstuff or Poisoning Incidents". In: *Genome announcements* 4.3, e00435–16. DOI: 10.1128/genomeA.00435-16.

Dai, Zhihao, Jean-Claude Sirard, Michele Mock, and Theresa M. Koehler (1995). "The *atxA* gene product activates transcription of the anthrax toxin genes and is essential for virulence". In: *Molecular Microbiology* 16.6, pp. 1171–1181. DOI: 10.1111/j.1365-2958.1995.tb02340.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-2958.1995.tb02340.x.

Dash, H.R., N. Mangwani, and S. Das (2014). "Characterization and potential application in mercury bioremediation of highly mercury-resistant marine bacterium *Bacillus thuringiensis* PW-05". In: *Environ Sci Pollut Res* 21, pp. 2642–2653. DOI: https://doi.org/10.1007/s11356-013-2206-8.

Doll, Etienne V., Siegfried Scherer, and Mareike Wenning (2017). "Spoilage of Microfiltered and Pasteurized Extended Shelf Life Milk Is Mainly Induced by Psychrotolerant Spore-Forming Bacteria that often Originate from Re-

contamination". In: *Frontiers in microbiology* 8, pp. 135–135. DOI: 10.3389/fmicb.2017.00135.

Drewnowska, Justyna M. and Izabela Swiecicka (2013). "Eco-Genetic Structure of *Bacillus cereus sensu lato* Populations from Different Environments in Northeastern Poland". In: *PLOS ONE* 8.12, pp. 1–11. DOI: 10.1371/journal.pone.0080175.

Duc, Le H., Huynh A. Hong, Teresa M. Barbosa, Adriano O. Henriques, and Simon M. Cutting (2004). "Characterization of *Bacillus* probiotics available for human use". In: *Applied and environmental microbiology* 70.4, pp. 2161–2171. DOI: 10.1128/aem.70.4.2161-2171.2004.

EFSA (2016). "Risks for public health related to the presence of *Bacillus cereus* and other *Bacillus* spp. including *Bacillus thuringiensis* in foodstuffs". In: *EFSA Journal* 14.7, e04524. DOI: 10.2903/j.efsa.2016.4524. eprint: https://efsa.onlinelibrary.wiley.com/doi/pdf/10.2903/j.efsa.2016.4524.

Ehling-Schulz, M. and U. Messelhausser (2013). "*Bacillus* next generation diagnostics: moving from detection toward subtyping and risk-related strain profiling". In: *Front Microbiol* 4, p. 32. DOI: 10.3389/fmicb.2013.00032.

Fox, G. E., J. D. Wisotzkey, and Jr. Jurtshuk P. (1992). "How close is close: 16S rRNA sequence identity may not be sufficient to guarantee species identity". In: *Int J Syst Bacteriol* 42.1, pp. 166–70. DOI: 10.1099/00207713-42-1-166.

Gee, Jay E., Chung K. Marston, Scott A. Sammons, Mark A. Burroughs, and Alex R. Hoffmaster (2014). "Draft Genome Sequence of *Bacillus cereus* Strain BcFL2013, a Clinical Isolate Similar to G9241". In: *Genome announcements* 2.3, e00469–14. DOI: 10.1128/genomeA.00469-14.

Guinebretiere, M. H., S. Auger, et al. (2013). "*Bacillus cytotoxicus* sp. nov. is a novel thermotolerant species of the *Bacillus cereus* Group occasionally associated with food poisoning". In: *Int J Syst Evol Microbiol* 63.Pt 1, pp. 31–40. DOI: 10.1099/ijs.0.030627-0.

Guinebretiere, M. H., F. L. Thompson, et al. (2008). "Ecological diversification in the *Bacillus cereus* Group". In: *Environ Microbiol* 10.4, pp. 851–65. DOI: 10.1111/j.1462-2920.2007.01495.x.

Guinebretiere, M. H., P. Velge, et al. (2010). "Ability of *Bacillus cereus* group strains to cause food poisoning varies according to phylogenetic affiliation (groups I to VII) rather than species affiliation". In: *J Clin Microbiol* 48.9, pp. 3388–91. DOI: 10.1128/JCM.00921-10.

Helgason, Erlendur, Nicolas J. Tourasse, Roger Meisal, Dominique A. Caugant, and Anne-Brit Kolsto (2004). "Multilocus sequence typing scheme for bacteria of the *Bacillus cereus* group". In: *Applied and environmental microbiology* 70.1, pp. 191–201. DOI: 10.1128/aem.70.1.191-201.2004.

Hendriksen, Niels Bohse, Bjarne Munk Hansen, and Jens Efsen Johansen (2006). "Occurrence and pathogenic potential of *Bacillus cereus* group bacteria in a sandy loam". In: 89, pp. 239–249. DOI: https://doi.org/10.1007/s10482-005-9025-y.

Hoffmaster, A. R. et al. (2008). "Genetic diversity of clinical isolates of *Bacillus cereus* using multilocus sequence typing". In: *BMC Microbiol* 8, p. 191. DOI: 10.1186/1471-2180-8-191.

Hoffmaster, Alex R. et al. (2004). "Identification of anthrax toxin genes in a *Bacillus cereus* associated with an illness resembling inhalation anthrax". In: *Proceedings of the National Academy of Sciences of the United States of America* 101.22, pp. 8449–8454. DOI: 10.1073/pnas.0402414101.

Hong, Huynh A., Le Hong Duc, and Simon M. Cutting (2005). "The use of bacterial spore formers as probiotics". In: *FEMS Microbiology Reviews* 29.4, pp. 813–835. DOI: 10.1016/j.femsre.2004.12.001. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1016/j.femsre.2004.12.001.

Hoton, F. M. et al. (2009). "Family portrait of *Bacillus cereus* and *Bacillus weihenstephanensis* cereulide-producing strains". In: *Environ Microbiol Rep* 1.3, pp. 177–83. DOI: 10.1111/j.1758-2229.2009.00028.x.

Huys, Geert et al. (2013). "Microbial characterization of probiotics–advisory report of the Working Group "8651 Probiotics" of the Belgian Superior Health Council (SHC)". In: *Molecular nutrition and food research* 57.8, pp. 1479–1504. DOI: 10.1002/mnfr.201300065.

Inouye, M. et al. (2014). "SRST2: Rapid genomic surveillance for public health and hospital microbiology labs". In: *Genome Med* 6.11, p. 90. DOI: 10.1186/s13073-014-0090-6.

Ivy, R. A. et al. (2012). "Identification and characterization of psychrotolerant sporeformers associated with fluid milk production and processing". In: *Appl Environ Microbiol* 78.6, pp. 1853–64. DOI: 10.1128/AEM.06536-11.

Jimenez, Guillermo, Anicet R. Blanch, Javier Tamames, and Ramon Rossello-Mora (2013). "Complete Genome Sequence of *Bacillus toyonensis* BCT-7112T, the Active Ingredient of the Feed Additive Preparation Toyocerin". In: *Genome announcements* 1.6, e01080–13. DOI: 10.1128/genomeA.01080-13.

Jimenez, G. et al. (2013). "Description of *Bacillus toyonensis* sp. nov., a novel species of the *Bacillus cereus* group, and pairwise genome comparisons of the species of the group by means of ANI calculations". In: *Syst Appl Microbiol* 36.6, pp. 383–91. DOI: 10.1016/j.syapm.2013.04.008.

Joensen, K. G. et al. (2014). "Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic *Escherichia coli*". In: *J Clin Microbiol* 52.5, pp. 1501–10. DOI: 10.1128/JCM.03617-13.

Johnson, Shannon L. et al. (2015). "Finished Genome Sequence of *Bacillus cereus* Strain 03BB87, a Clinical Isolate with *B. anthracis* Virulence Genes". In: *Genome announcements* 3.1, e01446–14. DOI: 10.1128/genomeA.01446-14.

Jouzani, G.S., E. Valijanian, and R. Sharafi (2017). "*Bacillus thuringiensis*: a successful insecticide with new environmental features and tidings". In: *Appl Microbiol Biotechnol* 101, pp. 2691–2711. DOI: 10.1007/s00253-017-8175-y.

Klee, S. R. et al. (2010). "The genome of a *Bacillus* isolate causing anthrax in chimpanzees combines chromosomal properties of *B. cereus* with *B. anthracis* virulence plasmids". In: *PLoS One* 5.7, e10986. DOI: 10.1371/journal.pone.0010986.

Ko, K. S. et al. (2003). "Identification of *Bacillus anthracis* by *rpoB* sequence analysis and multiplex PCR". In: *J Clin Microbiol* 41.7, pp. 2908–14.

Ko, Kwan Soo et al. (2004). "Population structure of the *Bacillus cereus* group as determined by sequence analysis of six housekeeping genes and the *plcR* Gene". In: *Infection and immunity* 72.9, pp. 5253–5261. DOI: 10.1128/IAI.72.9.5253-5261.2004.

Kodama, Y., M. Shumway, R. Leinonen, and Collaboration International Nucleotide Sequence Database (2012). "The Sequence Read Archive: explosive growth of sequencing data". In: *Nucleic Acids Res* 40.Database issue, pp. D54–6. DOI: 10.1093/nar/gkr854.

Kovac, J. et al. (2016). "Production of hemolysin BL by *Bacillus cereus* group isolates of dairy origin is associated with whole-genome phylogenetic clade". In: *BMC Genomics* 17, p. 581. DOI: 10.1186/s12864-016-2883-z.

Ladeuze, Sandy, Nathalie Lentz, Laurence Delbrassinne, Xiaomin Hu, and Jacques Mahillon (2011). "Antifungal Activity Displayed by Cereulide, the Emetic Toxin Produced by *Bacillus cereus*". In: *Applied and Environmental Microbiology* 77.7, pp. 2555–2558. DOI: 10.1128/AEM.02519-10. eprint: https://aem.asm.org/content/77/7/2555.full.pdf.

Lechner, S. et al. (1998). "*Bacillus weihenstephanensis* sp. nov. is a new psychrotolerant species of the *Bacillus cereus* group". In: *Int J Syst Bacteriol* 48 Pt 4, pp. 1373–82. DOI: 10.1099/00207713-48-4-1373.

Lee, Hyungjae, John J. Churey, and Randy W. Worobo (2009). "Biosynthesis and transcriptional analysis of thurincin H, a tandem repeated bacteriocin genetic locus, produced by *Bacillus thuringiensis* SF361". In: *FEMS Microbiology Letters* 299.2, pp. 205–213. DOI: 10.1111/j.1574-6968.2009.01749.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1574-6968.2009.01749.x.

Leinonen, R., H. Sugawara, M. Shumway, and Collaboration International Nucleotide Sequence Database (2011). "The sequence read archive". In: *Nucleic Acids Res* 39.Database issue, pp. D19–21. DOI: 10.1093/nar/gkq1019.

Lekota, Kgaugelo E. et al. (2015). "Draft Genome Sequences of Two South African *Bacillus anthracis* Strains". In: *Genome announcements* 3.6, e01313–15. DOI: 10.1128/genomeA.01313-15.

Liu, Y. et al. (2015a). "Genomic insights into the taxonomic status of the *Bacillus cereus* group". In: *Sci Rep* 5, p. 14082. DOI: 10.1038/srep14082.

— (2015b). "Genomic insights into the taxonomic status of the Bacillus cereus group". In: *Sci Rep* 5, p. 14082. DOI: 10.1038/srep14082.

Logan Niall A., Paul De Vos (2015). "*Bacillus*". In: *Bergey's Manual of Systematics of Archaea and Bacteria*. John Wiley and Sons, Inc., pp. 1–163. DOI: doi:10.1002/9781118960608.gbm00530.

Lucking, Genia, Marina Stoeckel, Zeynep Atamer, Jorg Hinrichs, and Monika Ehling-Schulz (2013). "Characterization of aerobic spore-forming bacteria associated with industrial dairy processing environments and product spoilage". In: *International Journal of Food Microbiology* 166.2, pp. 270–279.

Maechler, Martin, Peter Rousseeuw, Anja Struyf, Mia Hubert, and Kurt Hornik (2017). *cluster: Cluster Analysis Basics and Extensions*. R package version 2.0.6.

Martinez, Bismarck A., Jayne Stratton, and Andreia Bianchini (2017). "Isolation and genetic identification of spore-forming bacteria associated with concentrated-milk processing in Nebraska". In: *Journal of Dairy Science* 100.2, pp. 919–932. DOI: 10.3168/jds.2016-11660.

Miller, R. A., S. M. Beno, et al. (2016). "*Bacillus wiedmannii* sp. nov., a psychrotolerant and cytotoxic *Bacillus cereus* group species isolated from dairy foods and dairy environments". In: *Int J Syst Evol Microbiol* 66.11, pp. 4744–4753. DOI: 10.1099/ijsem.0.001421.

Miller, R. A., J. Jian, S. M. Beno, M. Wiedmann, and J. Kovac (2018). "Intraclade Variability in Toxin Production and Cytotoxicity of *Bacillus cereus* Group

Type Strains and Dairy-Associated Isolates". In: *Appl Environ Microbiol* 84.6. DOI: 10.1128/AEM.02479-17.

Miller, R. A., D. J. Kent, et al. (2015). "Spore populations among bulk tank raw milk and dairy powders are significantly different". In: *J Dairy Sci* 98.12, pp. 8492–504. DOI: 10.3168/jds.2015-9943.

Nakamura, L. K. (1998). "*Bacillus pseudomycoides* sp. nov". In: *Int J Syst Bacteriol* 48 Pt 3, pp. 1031–5. DOI: 10.1099/00207713-48-3-1031.

Oh, So-Young, Jonathan M. Budzik, Gabriella Garufi, and Olaf Schneewind (2011). "Two capsular polysaccharides enable *Bacillus cereus* G9241 to cause anthrax-like disease". In: *Molecular microbiology* 80.2, pp. 455–470. DOI: 10.1111/j.1365-2958.2011.07582.x.

Ohba, Michio, Eiichi Mizuki, and Akiko Uemori (2009). "Parasporin, a New Anticancer Protein Group from *Bacillus thuringiensis*". In: *Anticancer Research* 29.1, pp. 427–433. eprint: http://ar.iiarjournals.org/content/29/1/427.full.pdf+html.

Okinaka, Richard T. et al. (2014). "Genome Sequence of *Bacillus anthracis* STI, a Sterne-Like Georgian/Soviet Vaccine Strain". In: *Genome announcements* 2.5, e00853–14. DOI: 10.1128/genomeA.00853-14.

Oksanen, Jari et al. (2017). *vegan: Community Ecology Package*. R package version 2.4-2.

Pena-Gonzalez, Angela et al. (2017). "Draft Genome Sequence of *Bacillus cereus* LA2007, a Human-Pathogenic Isolate Harboring Anthrax-Like Plasmids". In: *Genome announcements* 5.16, e00181–17. DOI: 10.1128/genomeA.00181-17.

Price, Lance B. et al. (2012). "*Staphylococcus aureus* CC398: Host Adaptation and Emergence of Methicillin Resistance in Livestock". In: *mBio* 3.1. Ed. by Fernando Baquero. DOI: 10.1128/mBio.00305-11. eprint: https://mbio.asm.org/content/3/1/e00305-11.full.pdf.

Pruss, B. M., R. Dietrich, B. Nibler, E. Martlbauer, and S. Scherer (1999). "The hemolytic enterotoxin HBL is broadly distributed among species of the *Bacillus cereus* group". In: *Appl Environ Microbiol* 65.12, pp. 5436–42.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.

Rasko, David A., Michael R. Altherr, Cliff S. Han, and Jacques Ravel (2005). "Genomics of the *Bacillus cereus* group of organisms". In: *FEMS Microbiology Reviews* 29.2, pp. 303–329. DOI: 10.1016/j.fmrre.2004.12.005. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1016/j.fmrre.2004.12.005.

Rosenquist, H., L. Smidt, S. R. Andersen, G. B. Jensen, and A. Wilcks (2005). "Occurrence and significance of *Bacillus cereus* and *Bacillus thuringiensis* in ready-to-eat food". In: *FEMS Microbiol Lett* 250.1, pp. 129–36. DOI: 10.1016/j.femsle.2005.06.054.

Rossi-Tamisier, M., S. Benamar, D. Raoult, and P. E. Fournier (2015). "Cautionary tale of using 16S rRNA gene sequence similarity values in identification of human-associated bacterial species". In: *Int J Syst Evol Microbiol* 65.Pt 6, pp. 1929–34. DOI: 10.1099/ijs.0.000161.

Ruckert, Christian et al. (2012). "Draft Genome Sequence of *Bacillus anthracis* UR-1, Isolated from a German Heroin User". In: *Journal of Bacteriology* 194.21, pp. 5997–5998. DOI: 10.1128/JB.01410-12. eprint: https://jb.asm.org/content/194/21/5997.full.pdf.

Schmid, Daniela et al. (2016). "Elucidation of enterotoxigenic *Bacillus cereus* outbreaks in Austria by complementary epidemiological and microbiological investigations, 2013". In: *International Journal of Food Microbiology* 232, pp. 80–86.

Slowikowski, Kamil (2016). *ggrepel: Automatically Position Non-Overlapping Text Labels with 'ggplot2'*. R package version 0.6.5.

Sorokin, Alexei et al. (2006). "Multiple-locus sequence typing analysis of *Bacillus cereus* and *Bacillus thuringiensis* reveals separate clustering and a distinct

population structure of psychrotrophic strains". In: *Applied and environmental microbiology* 72.2, pp. 1569–1578. DOI: 10.1128/AEM.72.2.1569-1578.2006.

Stenfors Arnesen, L. P., A. Fagerlund, and P. E. Granum (2008). "From soil to gut: *Bacillus cereus* and its food poisoning toxins". In: *FEMS Microbiol Rev* 32.4, pp. 579–606. DOI: 10.1111/j.1574-6976.2008.00112.x.

Swiecicka, I. and P. De Vos (2003). "Properties of *Bacillus thuringiensis* isolated from bank voles". In: *Journal of Applied Microbiology* 94.1, pp. 60–64. DOI: 10.1046/j.1365-2672.2003.01790.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1046/j.1365-2672.2003.01790.x.

Swiecicka, Izabela, Geraldine A. Van der Auwera, and Jacques Mahillon (2006). "Hemolytic and Nonhemolytic Enterotoxin Genes are Broadly Distributed among *Bacillus thuringiensis* Isolated from Wild Mammals". In: *Microbial Ecology* 52, pp. 544–551. DOI: https://doi.org/10.1007/s00248-006-9122-0.

Takeno, Akira et al. (2012). "Complete genome sequence of *Bacillus cereus* NC7401, which produces high levels of the emetic toxin cereulide". In: *Journal of bacteriology* 194.17, pp. 4767–4768. DOI: 10.1128/JB.01015-12.

Tallent, S. M., K. M. Kotewicz, E. A. Strain, and R. W. Bennett (2012). "Efficient Isolation and Identification of *Bacillus cereus* Group". In: *Journal of Aoac International* 95.2, pp. 446–451. DOI: 10.5740/jaoacint.11-251.

Terzi, Britta von, Peter C. B. Turnbull, Steve E. Bellan, and Wolfgang Beyer (2014). "Failure of Sterne- and Pasteur-Like Strains of *Bacillus anthracis* to Replicate and Survive in the Urban Bluebottle Blow Fly *Calliphora vicina* under Laboratory Conditions". In: *PLOS ONE* 9.1, pp. 1–7. DOI: 10.1371/journal.pone.0083860.

Thomsen, M. C. et al. (2016). "A Bacterial Analysis Platform: An Integrated System for Analysing Bacterial Whole Genome Sequencing Data for Clinical Diagnostics and Surveillance". In: *PLoS One* 11.6, e0157718. DOI: 10.1371/journal.pone.0157718.

Thorsen, L. et al. (2006). "Characterization of emetic *Bacillus weihenstephanensis*, a new cereulide-producing bacterium". In: *Appl Environ Microbiol* 72.7, pp. 5118–21. DOI: 10.1128/AEM.00170-06.

Tourasse, Nicolas J. et al. (2011). "Extended and global phylogenetic view of the *Bacillus cereus* group population by combination of MLST, AFLP, and MLEE genotyping data". In: *Food Microbiology* 28.2, pp. 236–244.

Van der Auwera, Geraldine A., Michael Feldgarden, Roberto Kolter, and Jacques Mahillon (2013). "Whole-Genome Sequences of 94 Environmental Isolates of *Bacillus cereus Sensu Lato*". In: *Genome announcements* 1.5, e00380–13. DOI: 10.1128/genomeA.00380-13.

Vangay, P., E. B. Fugett, Q. Sun, and M. Wiedmann (2013). "Food microbe tracker: a web-based tool for storage and comparison of food-associated microbes". In: *J Food Prot* 76.2, pp. 283–94. DOI: 10.4315/0362-028X.JFP-12-276.

Wang, Gaoyan et al. (2014). "Bactericidal thurincin H causes unique morphological changes in *Bacillus cereus* F4552 without affecting membrane permeability". In: *FEMS Microbiology Letters* 357.1, pp. 69–76. DOI: 10.1111/1574-6968.12486. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1574-6968.12486.

Warda, Alicja K. et al. (2016). "Linking *Bacillus cereus* Genotypes and Carbohydrate Utilization Capacity". In: *PloS one* 11.6, e0156796–e0156796. DOI: 10.1371/journal.pone.0156796.

Wickham, Hadley (2009). *Ggplot2 : elegant graphics for data analysis*. Use R! New York: Springer, viii, 212 p.

— (2011). "The Split-Apply-Combine Strategy for Data Analysis". In: *2011* 40.1, p. 29. DOI: 10.18637/jss.v040.i01.

— (2017). *stringr: Simple, Consistent Wrappers for Common String Operations*. R package version 1.2.0.

Wickham, Hadley, Romain Francois, Lionel Henry, and Kirill Muller (2016). *dplyr: A Grammar of Data Manipulation*. R package version 0.5.0.

Wickham, Hadley, Jim Hester, and Romain Francois (2017). *readr: Read Rectangular Text Data*. R package version 1.1.0.

Yang, Yong, Hua Gu, et al. (2016). "Genotypic heterogeneity of emetic toxin producing *Bacillus cereus* isolates from China". In: *FEMS Microbiology Letters* 364.1. DOI: 10.1093/femsle/fnw237. eprint: http://oup.prod.sis.lan/femsle/article-pdf/364/1/fnw237/23928498/fnw237.pdf.

Yang, Yong, Xiaofeng Yu, et al. (2017). "Multilocus sequence type profiles of *Bacillus cereus* isolates from infant formula in China". In: *Food Microbiology* 62, pp. 46–50.

Zhang, S. et al. (2015). "*Salmonella* serotype determination utilizing high-throughput genome sequencing data". In: *J Clin Microbiol* 53.5, pp. 1685–92. DOI: 10.1128/JCM.00323-15.

Zhong, Wenwan, Yulin Shou, Thomas M. Yoshida, and Babetta L. Marrone (2007). "Differentiation of *Bacillus anthracis*, *B. cereus*, and *B. thuringiensis* by Using Pulsed-Field Gel Electrophoresis". In: *Applied and Environmental Microbiology* 73.10, pp. 3446–3449. DOI: 10.1128/AEM.02478-06. eprint: https://aem.asm.org/content/73/10/3446.full.pdf.

Zhu, Kui et al. (2016). "Probiotic *Bacillus cereus* Strains, a Potential Risk for Public Health in China". In: *Frontiers in microbiology* 7, pp. 718–718. DOI: 10.3389/fmicb.2016.00718.

Zwick, M. E. et al. (2012). "Genomic characterization of the *Bacillus cereus sensu lato* species: backdrop to the evolution of *Bacillus anthracis*". In: *Genome Res* 22.8, pp. 1512–24. DOI: 10.1101/gr.134437.111.

CHAPTER 5

CHARACTERIZATION OF EMETIC AND DIARRHEAL *BACILLUS CEREUS* STRAINS FROM A 2016 FOODBORNE OUTBREAK USING WHOLE-GENOME SEQUENCING: ADDRESSING THE MICROBIOLOGICAL, EPIDEMIOLOGICAL, AND BIOINFORMATIC CHALLENGES [1]

[1]FROM CARROLL, LAURA M., MARTIN WIEDMANN, MANJARI MUKHERJEE, DAVID C. NICHOLAS, LISA A. MINGLE, NELLIE B. DUMAS, JOCELYN A. COLE, AND JASNA KOVAC (2019). "CHARACTERIZATION OF EMETIC AND DIARRHEAL *BACILLUS CEREUS* STRAINS FROM A 2016 FOODBORNE OUTBREAK USING WHOLE-GENOME SEQUENCING: ADDRESSING THE MICROBIOLOGICAL, EPIDEMIOLOGICAL, AND BIOINFORMATIC CHALLENGES". IN: *FRONTIERS IN MICROBIOLOGY* 10, PP. 144. DOI: 10.3389/FMICB.2019.00144.

5.1 Abstract

The *Bacillus cereus* group comprises multiple species capable of causing emetic or diarrheal foodborne illness. Despite being responsible for tens of thousands of illnesses each year in the U.S. alone, whole-genome sequencing (WGS) is not yet routinely employed to characterize *B. cereus* group isolates from foodborne outbreaks. Here, we describe the first WGS-based characterization of isolates linked to an outbreak caused by members of the *B. cereus* group. In conjunction with a 2016 outbreak traced to a supplier of refried beans served by a fast food restaurant chain in upstate New York, a total of 33 *B. cereus* group isolates were obtained from human cases (n = 7) and food samples (n = 26). Emetic (n = 30) and diarrheal (n = 3) isolates were most closely related to *B. paranthracis* (group III) and *B. cereus sensu stricto* (group IV), respectively. WGS indicated that the 30 emetic isolates (24 and 6 from food and humans, respectively) were closely related and formed a well-supported clade distinct from publicly available emetic group III genomes with an identical sequence type (ST 26). The 30 emetic group III isolates from this outbreak differed from each other by a mean of 8.3 to 11.9 core single nucleotide polymorphisms (SNPs), while differing from publicly available emetic group III ST 26 *B. cereus* group genomes by a mean of 301.7 to 528.0 core SNPs, depending on the SNP calling methodology used. Using a WST-1 cell proliferation assay, the strains isolated from this outbreak had only mild detrimental effects on HeLa cell metabolic activity compared to reference diarrheal strain *B. cereus* ATCC 14579. We hypothesize that the outbreak was a single source outbreak caused by emetic group III *B. cereus* belonging to the *B. paranthracis* species, although food samples were not tested for presence of the emetic toxin cereulide. In addition to showcasing how WGS can be used

to characterize *B. cereus* group strains linked to a foodborne outbreak, we also discuss potential microbiological and epidemiological challenges presented by *B. cereus* group outbreaks, and we offer recommendations for analyzing WGS data from the isolates associated with them.

5.2 Introduction

The *Bacillus cereus* (*B. cereus*) group, also known as *B. cereus sensu lato* (*s.l.*) is a complex of closely related species that vary in their ability to cause disease in humans. Foodborne illness caused by members of the group primarily manifests itself in one of two forms: (i) emetic intoxication that is caused by cereulide, a heat-stable toxin produced by *B. cereus* within a food matrix prior to consumption, or (ii) a diarrheal toxicoinfection, caused by enterotoxins produced by bacteria in the small intestine of the host (Ehling-Schulz, Fricker, and Scherer 2004; Schoeni and Wong 2005; Stenfors Arnesen, Fagerlund, and Granum 2008). Here we refer to isolates that carry *ces* genes encoding the cereulide biosynthetic pathway as emetic isolates, and isolates that lack *ces* genes but carry either *hbl* or *cytK-2* genes that encode diarrheal enterotoxins as diarrheal isolates. The gene variant *cytK-2* was included in this definition, as it was previously found in non-*B. cytotoxicus* isolates associated with diarrheal illness (Castiaux et al. 2015; Miller, Jian, et al. 2018). The presence of *nhe* genes was not included in our present definition of diarrheal isolates, due to the fact that *nhe* genes are ubiquitously found in the majority of the *B. cereus* group population (Carroll et al. 2017; Miller, Jian, et al. 2018), including all isolates in the present study, and their contribution to diarrheal toxicoinfection is not yet fully understood (Doll, Ehling-Schulz, and Vogelmann 2013).

As foodborne pathogens, members of the *B. cereus* group are estimated to cause 63,400 foodborne disease cases per year in the U.S. (Scallan et al. 2011) and are confirmed or suspected to have been responsible for 235 outbreaks reported in the U.S. between 1998 and 2008 (Bennett, K. A. Walsh, and Gould 2013). Due in part to its typically self-limiting nature, foodborne illness caused by members of the *B. cereus* group is under-reported (Granum and Lund 1997; Stenfors Arnesen, Fagerlund, and Granum 2008), although severe infections resulting in patient death have been reported (Naranjo et al. 2011; Sanaei-Zadeh 2012; Lotte et al. 2017). Furthermore, *B. cereus* group isolates that have been linked to human clinical cases of foodborne disease rarely undergo whole-genome sequencing (WGS), as is becoming the norm for other foodborne pathogens (Joensen et al. 2014; Ashton et al. 2015; Moura et al. 2017).

Here, we describe a foodborne outbreak caused by members of the *B. cereus* group in which WGS was implemented to characterize isolates from human clinical cases and food. To our knowledge, this is the first description of a *B. cereus* outbreak in which WGS was employed to characterize isolates. By testing various combinations of variant calling methodologies, we showcase how different bioinformatics pipelines can yield vastly different results when pairwise SNP differences are the desired metric for determining whether an isolate is part of an outbreak or not. In addition to discussing the bioinformatic challenges, we examine potential microbiological and epidemiological obstacles that can hinder characterization of *B. cereus* group isolates from suspected foodborne outbreaks, and we offer recommendations to guide the characterization of future *B. cereus* group outbreaks using WGS.

5.3 Materials and Methods

5.3.1 Collection of Epidemiological Data

Epidemiological investigations were coordinated by the New York State Department of Health (NYSDOH), and the outbreak was reported to the U.S. Centers for Disease Control and Prevention (CDC). Investigation methods included (i) a cohort study, (ii) food preparation review, (iii) an investigation at a factory/production/treatment plant, (iv) food product traceback, and (v) environment/food/water sample testing.

5.3.2 Isolation and Initial Characterization of *B. cereus* Strains

Stool specimens were plated directly onto mannitol-egg yolk-polymyxin (MYP) agar and incubated aerobically at 37°C for 24 h. Food samples were diluted 1:10 in 1 X PBS, pH 7.4 in a filter bag for homogenizer blenders and homogenized for 2 min. One hundred μl of each homogenized sample were plated onto MYP agar and incubated aerobically at 37°C for 24 h. The MYP agar plates for both the stool specimens and food samples were observed after the 24 h incubation period. Individual *B. cereus*-like colonies (i.e., pink colored and lecithinase positive) were subcultured on trypticase soy agar (TSA) plates supplemented with 5% sheep blood and incubated aerobically at 37°C for 18-24 h. These isolates were identified as *B. cereus* using the following conventional microbiological techniques: Gram stain, colony morphology, hemolysis, motility, and spore stain. To test for the presence of parasporal crystals often associated with *B. thuringiensis*, isolates were cultured for 48 h at 37°C on sporulation agar slants.

Smears were prepared, and slides were heat fixed and then stained using malachite green and counter stained with carbol fuchsin (Tallent, Rhodehamel, et al. 1998). Slides were then observed for the presence or absence of parasporal crystals.

5.3.3 *rpoB* Allelic Typing

The 33 outbreak isolates were streaked onto brain heart infusion (BHI) agar from their respective cryo stocks stored at −80°C and incubated overnight at 37°C. Single isolated colonies were inoculated in 5 ml BHI broth and incubated overnight at 32°C and used for genomic DNA extraction using Qiagen DNeasy blood and tissue kits (Qiagen). Extracted DNA was used as a template in a PCR reaction using primers targeting a 750 bp sequence of the *rpoB* gene (RzrpoBF: AARYTIGGMCCTGAAGAAAT and RZrpoBR: TGIARTTRTCATCAACCATGTG) (Ivy et al. 2012). PCR was carried out in 25 μl reactions using GoTaq Green Master Mix (Promega Corporation) under the following thermal cycling conditions: 3 min at 94°C, followed by 40 cycles of 30 s at 94°C, 30 s at 55 – 45°C (in the first 20 cycles, the temperature was reduced for 0.5°C per cycle and then kept at 45°C in the following 20 cycles), followed by 1 min at 72°C, and a final hold at 4°C. The resulting PCR product was used for genotyping and preliminary species identification using the *rpoB* allele type database available in Food Microbe Tracker (Ivy et al. 2012; Vangay et al. 2013).

5.3.4 Bacterial Growth Conditions and Collection of Bacterial Supernatants

The 33 outbreak isolates, as well as *B. cereus s.s.* type strain ATCC 14579 and *B. cereus* emetic reference strain DSM 4312 (Food Microbe Tracker ID FSL M8-0547) (Vangay et al. 2013) were streaked onto BHI agar from their respective cryo stocks stored at −80°C. For immunoassays and cytotoxicity assays (see sections "Hemolysin BL and Non-hemolytic Enterotoxin Detection" and "WST-1 Metabolic Activity Assay"), cultures grown from single isolated colonies for 18 h at 37°C without shaking were used for inoculation of fresh BHI broth. Fresh cultures were grown to early stationary phase as determined by an OD600 of ~ 1.5, which equals ~ 10^8 CFU/ml. After incubation, growth was quenched by placing cultures on ice. The cultures were then spun down at 16,000 g for 2 min, and the supernatants were collected, aliquoted in duplicate, and stored at −80°C until further use in cytotoxicity assays.

5.3.5 Hemolysin BL and Non-hemolytic Enterotoxin Detection

Diarrheal strains grown as described above were used for qualitative detection of hemolysin BL (Hbl) and non-hemolytic enterotoxins (Nhe) with the Duopath Cereus Enterotoxins immunoassay (Merck). Only select representatives of emetic outbreak strains were tested (i.e., FSL R9-6381, FSL R9-6382, FSL R9-6384, FSL R9-6389, FSL R9-6395, and FSL R9-6399), as they did not carry genes encoding Hbl and were therefore not expected to produce Hbl. Briefly, the temperatures of the cultures and immunoassay kits were adjusted to room temperature. 150 μl of each isolate culture were added to the immunoassay port,

following the manufacturer's instructions. The results were read as positive if a red test line was visible after a 30-min incubation at room temperature. Tests were considered valid only when control lines were visible.

5.3.6 WST-1 Metabolic Activity Assay

HeLa cells were seeded in 96-well plates at a seeding density of $8 \times 10^4 cells/cm^2$ (Fisichella et al. 2009) in Eagle's minimum essential medium (EMEM) supplemented with 10% fetal bovine serum (FBS) and allowed to grow for 18-24 h at 37°C, 5% CO_2. After incubation, the medium in each well was replaced with 100 μl of fresh medium containing 5% v/v of bacterial supernatants (prepared as described in section "Bacterial Growth Conditions and Collection of Bacterial Supernatants") that were thawed and pre-warmed to 37°C. The combined medium and supernatants were added to the cells using a multichannel pipettor to minimize the variability in the duration of cell exposure to the toxin amongst wells of a 96-well plate. Medium containing 5% BHI was used as a negative control. Medium containing 5% v/v of 1% Triton X-100 dissolved in BHI (final concentration in the test well was 0.05%) was used as a positive control expected to significantly reduce the viability of HeLa cells. After 15 min of intoxication at 37°C, 5% CO_2 (Miller, Jian, et al. 2018), 10 μl of WST-1 dye solution (Roche) was added to each well of the plate, and the plate was incubated for 25 min at 37°C, 5% CO_2, resulting in a total of 40 min exposure of cells to the supernatants. After 30 s of orbital shaking at 600 rpm, the absorbances were read by a microplate reader (Thermo Scientific Multiskan GO, Thermo Fisher Scientific) in a precision mode at 450 and 690 nm, the latter being subtracted from the former to account for the background signal (i.e., corrected absorbances) (Fisichella et al.

2009). Each test, including 0.05% Triton X-100, was conducted with six technical replicates and on two different HeLa passages using supernatants from single biological replicates, resulting in a total of 12 technical replicates per isolate. The viability of cells was determined by calculating a ratio of corrected absorbances to that of BHI, converting to percentages, and calculating the mean of the technical replicates for each isolate. The results were compared to the results for cells treated with (i) 0.05% Triton X-100, (ii) *B. cereus s.s.* type strain ATCC 14579 supernatant (i.e., reference for diarrheal strains), and (iii) *B. cereus* group strain DSM 4312 supernatant (i.e., reference for emetic strains).

5.3.7 Statistical Analysis of Cytotoxicity Data

A Welch's test and the Games-Howell post-hoc test that are appropriate for analyses of data with non-homogeneous variances were performed using results of all 12 technical replicates of each outbreak-associated isolate, as well as the reference strains and the positive control. For the Games-Howell test, a Bonferroni correction was applied to correct for multiple comparisons. Statistical analyses were carried out in R version 3.4.3 (R Core Team 2018).

5.3.8 Whole-Genome Sequencing

Genomic DNA was extracted from overnight cultures (~ 18 h) grown in BHI at 32°C using Qiagen DNeasy blood and tissue kits (Qiagen) or the Omega E.Z.N.A. Bacterial DNA kit (Omega) following the manufacturers' instructions. For the E.Z.N.A. Bacterial DNA kit, the additional steps recommended for

difficult-to-lyse bacteria were taken to obtain sufficient DNA yield. Briefly, one ml of an overnight culture was additionally treated with glass beads provided in the E.Z.N.A. kit. DNA was quantified using Qubit 3 and used for Nextera XT library preparation (Illumina). Pooled libraries were sequenced in two Illumina sequencing runs with either 2 x 250 or 2 x 300 bp reads at the Penn State Genomics Core Facility and at the Cornell Animal Health Diagnostic Center.

5.3.9 Initial Data Processing and Genome Assembly

Illumina adapters and low-quality bases were trimmed using Trimmomatic version 0.36 (Bolger, Lohse, and Usadel 2014) and the default parameters for Nextera paired-end reads, and FastQC version 0.11.5 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) was used to confirm that read quality was adequate (e.g., no reads flagged as poor quality, no Illumina adapters present). Genomes listed in Supplementary Table S1 were assembled *de novo* using SPAdes version 3.11.0 (Bankevich et al. 2012), and average per-base coverage was calculated using Samtools version 1.6 (H. Li, Handsaker, et al. 2009) after mapping reads to their respective *de novo* assemblies using BWA MEM version 0.7.13 (default parameters) (H. Li and Durbin 2010).

5.3.10 *In silico* Typing and Virulence Gene Detection

BTyper version 2.2.0 (Carroll et al. 2017) was used to perform *in silico* virulence gene detection, multi-locus sequence typing (MLST), *panC* group assignment (as

defined by Guinebretiere et al., 2010), and *rpoB* allelic typing, as well as to extract the gene sequences for all detected loci (Guinebretiere, Velge, et al. 2010). For virulence gene detection, the default settings were used (i.e., 50% amino acid sequence identity, 70% query coverage), as these cut-offs have been shown to correlate with PCR-based detection of virulence genes in *B. cereus* group isolates (J. Kovac et al. 2016; Carroll et al. 2017). BMiner version 2.0.2 (Carroll et al. 2017) was used to aggregate the output files from BTyper and create a virulence gene presence/absence matrix.

5.3.11 Construction of *k*-mer Based Phylogeny Using Outbreak Strains and Genomes of 18 *B. cereus* Group Species

kSNP version 3.1 (Gardner and Hall 2013; Gardner, Slezak, and Hall 2015) was used to produce a set of core SNPs among the 33 outbreak genomes, plus a type strain or RefSeq reference genome assembly from each of the 18 *B. cereus* group species listed in Supplementary Table S2 (Stenfors Arnesen, Fagerlund, and Granum 2008; Guinebretiere, Auger, et al. 2013; Jimenez et al. 2013; Miller, Beno, et al. 2016; Liu et al. 2017), using the optimal *k*-mer size as determined by Kchooser (k = 21). The resulting core SNPs were used in conjunction with RAxML version 8.2.11 (Stamatakis 2014) to construct a maximum likelihood (ML) phylogeny using the GTRCAT model with a Lewis ascertainment bias correction (Lewis 2001) to account for the use of solely variant sites, and 500 bootstrap replicates. The resulting phylogenetic tree was formatted using the phylobase (R Hackathon et al. 2019), ggtree (Yu et al. 2017), phytools (Revell 2012), and ape (Paradis, Claude, and Strimmer 2004) packages in R version

3.4.3.

5.3.12 Variant Calling and Phylogeny Construction Using Outbreak Isolates

Combinations of five reference-based variant calling pipelines (Table 5.1) and reference genomes (Table 5.2), as well as one reference-free SNP calling pipeline (Table 5.1), were used to separately identify core and total SNPs among (i) all 33 outbreak-related isolates (30 emetic group III isolates and three group IV isolates) and (ii) the subset of 30 emetic group III isolates. For the subset of 30 emetic group III isolates, all reference-based variant calling pipelines described below were additionally run with dustmasked versions of the reference genomes listed in Table 5.2, in which DustMasker version 1.0.0 (part of BLAST version 2.6.0) (Morgulis et al. 2006) was used to mask low-complexity portions (i.e., intervals with highly biased nucleotide distributions which can bias sequence similarity searches) in each reference genome (Ye, McGinnis, and Madden 2006).

Table 5.1: Description of variant calling pipelines and associated input data formats tested in this study.

Pipeline[a]	*Approach*	*Reference-based*	*Input data (file format)*[b]	*Read mapper*	*Variant caller*	*Reference(s) and in-depth pipeline descriptions*
CFSAN	Read mapping	Yes	PE reads (fastq)	Bowtie2	Varscan	https://snp-pipeline.readthedocs.io/en/latest/
Freebayes	Read mapping	Yes	PE reads (fastq)	BWA MEM	Freebayes	https://github.com/lmc297/SNPBac
kSNP3	*k*-mer based	No	Contigs (fasta)	Not applicable	kSNP3	https://sourceforge.net/projects/ksnp/files/
LYVE-SET	Read mapping	Yes	PE reads (fastq)	SMALT	Varscan	https://github.com/lskatz/lyve-SET
Parsnp	Core genome alignment	Yes	Contigs (fasta)	Not applicable	Parsnp	https://harvest.readthedocs.io/en/latest/content/parsnp.html
Samtools	Read mapping	Yes	PE reads (fastq)	BWA MEM	Samtools/ Bcftools	https://github.com/lmc297/SNPBac

[a]CFSAN, U.S. Food and Drug Administration (FDA) Center for Food Safety and Applied Nutrition SNP pipeline; LYVE-SET, U.S. Centers for Disease Control and Prevention (CDC) *Listeria*, *Yersinia*, *Vibrio*, and *Enterobacteriaceae* SNP Extraction Tool

[b]PE reads, Illumina paired-end reads

Table 5.2: Reference genomes used for reference-based variant calling in this study.

Reference genome	*Phylogenetic group*[a]	*Data set(s)*[b]	*ANI range*[c]	*NCBI accession number*	*Assembly level*	*Rationale for selection*
B. cereus strain ATCC 14579 chromosome	IV	All 33 isolates from two clades (clades III and IV)	98.8-98.9 (clade IV) 91.8-92.3 (clade III)	NC_004722.1	Complete Genome	*B. cereus s.s.* type strain; RefSeq reference genome; member of *panC* clade IV, the same clade as the three non-emetic outbreak-associated isolates sequenced in this study
B. cereus strain AH187 chromosome	III	All 33 isolates from two clades (clades III and IV); 30 emetic clade III isolates	92.0-92.2 (clade IV) 99.8-99.9 (clade III)	NC_011658.1	Complete Genome	Human clinical isolate associated with an emetic outbreak in 1972 (cooked rice, United Kingdom); identical virulotype, MLST sequence type, *rpoB* allelic type, and *panC* clade as 30 emetic outbreak isolates sequenced in this study
B. cytotoxicus strain NVH 391-98 chromosome	VII	All 33 isolates from two clades (clades III and IV)	82.6-82.7 (clade IV) 82.5-82.9 (clade III)	NC_009674.1	Complete Genome	Type strain of *B. cytotoxicus*, the most distant member of the *B. cereus* group as currently defined; shares a common ancestor with all isolates sequenced in this study
FOOD_10_19_16_RSNT1_2H_R9-6393	III	30 emetic clade III isolates	92.0-92.2 (clade IV) 100[d]-100 (clade III)	SRR6825038	Contigs	Emetic isolate from the outbreak reported here; assembly had high per-base coverage, as well as the fewest number of contigs of all genome assemblies from isolates in this outbreak

[a]Group determined via *panC* clade assignment function in BTyper version 2.2.0
[b]Data set(s) in this study for which a given genome was used as a reference genome for reference-based SNP calling
[c]Minimum and maximum average nucleotide identity (ANI) values of reference strain relative to clade IV and clade III genomes sequenced in this outbreak (n = 3 and 30, respectively) calculated using FastANI
[d]Minimum ANI value was less than 100 prior to rounding

For the Samtools and Freebayes pipelines (Table 5.1), trimmed Illumina paired-end reads from the queried isolates were mapped to the appropriate reference genome using BWA mem version 0.7.13 (Heng Li 2013) and either Samtools/Bcftools version 1.6 (H. Li, Handsaker, et al. 2009) or Freebayes version 1.1.0 (Garrison and Marth 2012), respectively, were used to call variants. Vcftools version 0.1.14 (Danecek et al. 2011) was used to remove indels and SNPs with a SNP quality score < 20, as well as to construct consensus sequences. For both variant calling pipelines, Gubbins version 2.2.0 (Croucher et al. 2015) was used to remove recombination events from the consensus sequences, and the Neighbor Similarity Score (NSS) (Jakobsen and Easteal 1996), Maximum Chi-Squared (Smith 1992), and Pairwise Homoplasy Index (PHI) (Bruen, Philippe, and Bryant 2006) tests implemented in PhiPack version 1.0 (Bruen, Philippe, and Bryant 2006) were used to assess whether recombination and homoplasies were present in sequence alignments before and

after recombination was removed, using 1,000 permutations each and a window size of 100 (Supplementary Table S3). Both of these pipelines are publicly available and can be reproduced in their entirety (SNPBac version 1.0.0; https://github.com/lmc297/SNPBac).

For the CFSAN (Davis et al. 2015) and LYVE-SET (Katz et al. 2017) pipelines (versions 1.0.1 and 1.1.4 g, respectively; Table 5.1), trimmed Illumina paired-end reads were used as input, and all default pipeline steps were run as outlined in the manuals. For the Parsnp pipeline (Treangen et al. 2014) (Table 5.1), assembled genomes of the outbreak isolates were used as input, and Parsnp's implementation of PhiPack (Bruen, Philippe, and Bryant 2006) was used to filter out recombination events. For kSNP3 (Table 5.1), assembled genomes of the outbreak isolates were used as input, and Kchooser was used to determine the optimum *k*-mer size for the full 33-isolate data set and the 30 emetic group III isolate set (k = 21 and 23, respectively).

For all variant calling and filtering pipelines, RAxML version 8.2.10 was used to construct ML phylogenies using the resulting SNPs under the GTRGAMMA model with a Lewis ascertainment bias correction and 1,000 bootstrap replicates. Phylogenetic trees were annotated using FigTree version 1.4.3 (http://tree.bio.ed.ac.uk/software/figtree/).

5.3.13 Variant Calling and Statistical Comparison of Emetic Outbreak Isolates to Publicly Available Genomes

To compare emetic group III isolates from this outbreak to other emetic group III isolates, BTyper version 2.2.1 was used to query all 2,156 *B. cereus* group genome assemblies available in NCBI's RefSeq database (downloaded March 2018) (Pruitt, Tatusova, and Maglott 2007) and identify all genome assemblies that (i) belonged to group III based on *panC* sequence, (ii) belonged to ST 26 based on *in silico* MLST, and (iii) were found to possess the *ces* operon in its entirety (*cesABCD*) at the default coverage and identity thresholds. This search produced 25 genome assemblies in addition to the 30 emetic group III genomes sequenced here. Only three of the 25 RefSeq genome assemblies had Sequence Read Archive (SRA) data linked to their BioSample accession numbers, making short read data readily available only for these three isolates. Consequently, only Parsnp version 1.2 and kSNP version 3.1 were used to identify SNPs in all 55 group III emetic genomes (25 from NCBI RefSeq and 30 sequenced here), as these approaches can be used with assembled genomes and do not require short reads as input. For Parsnp, the chromosome of *B. cereus* AH187 was used as a reference genome. For kSNP3, Kchooser was used to select the optimal *k*-mer size ($k = 21$), and the chromosome of *B. cereus* AH187 was included for *k*-mer based SNP calling.

RAxML version 8.2.10 was used to construct ML phylogenies using the resulting core SNPs for each of the Parsnp and kSNP3 pipelines under the GTR-CAT model with a Lewis ascertainment bias correction and 1,000 bootstrap replicates. Pairwise core SNP differences between all 55 isolates were obtained using the dist.gene function in R's ape package. The permutest and betadisper

functions in R's vegan package (Oksanen et al. 2017) were used to conduct an ANOVA-like permutation test to test if publicly available genomes were more variable than isolates from this outbreak based on pairwise core SNP differences and 5 independent trials using 100,000 permutations each. Analysis of similarity (ANOSIM) using the anosim function in the vegan package in R was used to determine if the average of the ranks of within-group distances was greater than or equal to the average of the ranks of between-group distances (Clarke 1993; Anderson and D. C. I. Walsh 2013), where groups were defined as (i) the 30 emetic isolates from this outbreak, and (ii) the 25 external emetic ST 26 isolates (downloaded from RefSeq). ANOSIM tests were conducted using pairwise core SNP differences and five independent runs of 10,000 permutations each. For both the ANOVA-like permutation tests and the ANOSIM tests, Bonferroni corrections were used to correct for multiple comparisons at the $\alpha = 0.05$ level.

5.3.14 Statistical Comparison of Phylogenetic Trees

The Kendall-Colijn (Kendall and Colijn 2015; Kendall and Colijn 2016) test described by Katz et al. (Katz et al. 2017) was used to compare the topologies of trees, using the treespace (Jombart et al. 2017), ips (Heibl 2008 onwards), phangorn (Schliep et al. 2017), docopt (de Jonge 2018), and stringr (Wickham 2017) packages in R version 3.4.3. The phylogenies that underwent pairwise testing were constructed using (i) either core or total SNPs identified in 30 emetic group III genomes via all six SNP calling pipelines (Table 5.1), using either an unmasked or dustmasked closed reference genome (*B. cereus* AH187; Table 5.2), and (ii) SNPs identified in 55 emetic ST 26 genomes (25 publicly available genomes and the 30 emetic isolates sequenced here) using the kSNP3 (core

and total SNPs) and Parsnp (core SNPs, as Parsnp queries the core genome by definition) pipelines. For all pairwise tree comparisons, a lambda value of 0 (to give weight to tree topology rather than branch lengths) (Katz et al. 2017) was used along with 100,000 random trees as a background distribution, and a Bonferroni correction was used to correct for multiple comparisons. Pairs of trees were considered to be more topologically similar than would be expected by chance if a significant *P*-value ($P < 0.05$) resulted after correcting for multiple testing (Katz et al. 2017).

5.3.15 Calculation of Average Nucleotide Identity Values

FastANI version 1.0 (Jain et al. 2018) was used to calculate average nucleotide identity (ANI) values between assembled genomes of isolates sequenced in this study and selected reference genomes (Table 5.2), as well as the genomes of 18 currently published *B. cereus* group species (Supplementary Table S2).

5.3.16 Supplementary Material and Availability of Data

Trimmed Illumina reads for all 33 isolates sequenced in this study have been made publicly available (NCBI BioProject Accession PRJNA437714), with NCBI BioSample and SRA accession numbers for all isolates listed in Supplementary Table S1. All figures have been deposited in FigShare (DOI https://doi.org/10.6084/m9.figshare.7001525.v1), and records of all isolates are available in Food Microbe Tracker (Vangay et al. 2013).

5.4 Results

5.4.1 Both Emetic and Diarrheal Symptoms Were Reported Among Cases Associated With the *B. cereus* Foodborne Outbreak

Between September 30th and October 6th, 2016, local health departments in upstate New York's Niagara and Erie counties reported a total of 179 estimated foodborne illness cases among customers of a Mexican fast-food restaurant chain in eight towns/cities. Among these cases, laboratory results were available for ten cases. For seven of these cases, *B. cereus* group species were isolated from patient stool samples. While no deaths, hospitalizations, or emergency room visits were reported from 169 cases from which information was obtained, 4 resulted in a visit to a health care provider (not including emergency room visits). More than 2/3 of 179 cases were female (69%), and 61% of cases fell within the 20-74 age group. In 156 of 179 total cases (87%), refried beans had been consumed.

Of 169 cases from which information was obtained, 88% reported vomiting, and more than half reported nausea and abdominal cramps (95 and 65%, respectively). However, in addition to vomiting, 38% of cases also reported diarrhea. Additional symptoms reported included (i) weakness (43%), (ii) chills (40%), (iii) dehydration (35%), (iv) headache (28%), (v) myalgia (muscle ache/pain; 16%), (vi) fever (16%), (vii) sweating (16%), and (viii) sore throat (3%). The incubation period observed for all cases ranged from 0.25 to 24 h, with a median of 2 h. The duration of illness ranged from 0.25 to 144 h, with a median estimate

of 6 h.

A traceback was conducted, with the source of the outbreak determined to be a processing plant in Pennsylvania. The distributor in Pennsylvania packaged the refried beans specifically for the chain establishment where the outbreak occurred. The establishments where the outbreak occurred received 5 lb trays of pre-cooked, sealed, and frozen refried beans from the production/packaging facility. The refried beans would undergo cooking and a hot hold prior to consumption at the establishments where the outbreak occurred. It was determined that the refried beans were contaminated prior to preparation at the chain establishment.

Stool samples from suspect cases were cultured on MYP agar and *B. cereus*-like colonies were isolated from seven stool samples. Additionally, *B. cereus*-like colonies were isolated from nine food samples that were collected from five restaurants. In total, seven isolates from stool samples and 26 isolates from foods were confirmed to belong to the *B. cereus* group using standard microbiological methods. Isolates that were large Gram-positive rods, beta-hemolytic, and motile were presumptively identified as *B. cereus*-like. Additionally, spore staining was performed to test for the presence of parasporal crystals associated with *B. thuringiensis*, for which all isolates were negative. All 33 *B. cereus* group isolates underwent preliminary molecular characterization by Sanger sequencing of *rpoB*, which revealed two distinct allelic types belonging to phylogenetic groups III (*rpoB* allelic type AT 125) and IV (AT 92).

5.4.2 WGS Confirms Presence of Multiple *B. cereus* Group Species Represented Among Outbreak Strains

rpoB allelic types (ATs) assigned *in silico* were identical to those obtained using Sanger sequencing for all 33 isolates (Table 5.3). *panC* group assignment confirmed the presence of *B. cereus s.l.* isolates from multiple phylogenetic groups (Table 5.3), with *panC* group III (n = 30) and *panC* group IV (n = 3) represented among the 33 isolates. *In silico* MLST further resolved the group IV isolates into two sequence types (STs): the two strains isolated from refried beans served at two different restaurants had identical STs, while the single human isolate belonging to group IV had a unique ST (Table 5.3). All 30 *panC* group III isolates belonged to ST 26, including the remaining six human clinical isolates (Table 5.3).

The presence of isolates from multiple *B. cereus s.l.* phylogenetic groups, as suggested by the *rpoB*, *panC*, and MLST loci among isolates sequenced in conjunction with this outbreak, was confirmed using core SNPs detected in all outbreak isolates, as well as the genomes of 18 currently recognized *B. cereus* group species (Figure 5.1). The three isolates assigned to *panC* group IV using a 7-group scheme (Guinebretiere, Thompson, et al. 2008) were most closely related to the *B. cereus s.s.* type strain (Figure 5.1). All three group IV *B. cereus* isolates possessed diarrheal toxin genes *hblABCD* and *cytK-2* at high identity and coverage (Figure 5.1), which code for enterotoxins hemolysin BL (Hbl) and cytotoxin K variant 2 (CytK-2), respectively. The 30 isolates assigned to *panC* group III, however, were most closely related to the type strain of *B. paranthracis* (Figure 5.1). Unlike the *B. paranthracis* type strain, all of the group III isolates investigated here were motile and possessed the *cesABCD* operon (Figure 5.1),

Table 5.3: List of outbreak isolates and corresponding metadata, single- and multi-locus sequence types, and species.

Isolate name	*Source (General)*	*Source (Specific)*	*Isolation date*	*Production Date/Batch*[a]	*panC Group*[b]	*MLST ST*[c]	*rpoB AT*[d]	*Closest Type Strain (ANI)*[e]
FOOD_10_18_16_LFTOV_NA_R9-6400	Food	Leftovers	18-Oct	Unknown	III	26	125	*B. paranthracis* MN5 (97.5)
FOOD_10_18_16_LFTOV_NA_R9-6401	Food	Leftovers	18-Oct	Unknown	III	26	125	*B. paranthracis* MN5 (97.5)
FOOD_10_18_16_LFTOV_NA_R9-6402	Food	Leftovers	18-Oct	Unknown	III	26	125	*B. paranthracis* MN5 (97.5)
FOOD_10_19_16_RSNT1_1B_R9-6388	Food	Restaurant 1	19-Oct	1/B	III	26	125	*B. paranthracis* MN5 (97.5)
FOOD_10_19_16_RSNT1_1B_R9-6389	Food	Restaurant 1	19-Oct	1/B	III	26	125	*B. paranthracis* MN5 (97.5)
FOOD_10_19_16_RSNT1_1B_R9-6390	Food	Restaurant 1	19-Oct	1/B	III	26	125	*B. paranthracis* MN5 (97.5)
FOOD_10_19_16_RSNT1_1B_R9-6391	Food	Restaurant 1	19-Oct	1/B	III	26	125	*B. paranthracis* MN5 (97.5)
FOOD_10_19_16_RSNT1_2A_R9-6386	Food	Restaurant 1	19-Oct	2/A	III	26	125	*B. paranthracis* MN5 (97.5)
FOOD_10_19_16_RSNT1_2A_R9-6387	Food	Restaurant 1	19-Oct	2/A	III	26	125	*B. paranthracis* MN5 (97.5)
FOOD_10_19_16_RSNT1_2H_R9-6392	Food	Restaurant 1	19-Oct	2/H	III	26	125	*B. paranthracis* MN5 (97.5)
FOOD_10_19_16_RSNT1_2H_R9-6393	Food	Restaurant 1	19-Oct	2/H	III	26	125	*B. paranthracis* MN5 (97.5)
FOOD_10_19_16_RSNT1_2H_R9-6394	Food	Restaurant 1	19-Oct	2/H	III	26	125	*B. paranthracis* MN5 (97.5)
FOOD_10_19_16_RSNT1_2H_R9-6395	Food	Restaurant 1	19-Oct	2/H	III	26	125	*B. paranthracis* MN5 (97.5)
FOOD_10_19_16_RSNT1_2H_R9-6396	Food	Restaurant 1	19-Oct	2/H	III	26	125	*B. paranthracis* MN5 (97.5)
FOOD_10_19_16_RSNT2_2A_R9-6397	Food	Restaurant 2	19-Oct	2/A	III	26	125	*B. paranthracis* MN5 (97.5)
FOOD_10_19_16_RSNT2_2A_R9-6398	Food	Restaurant 2	19-Oct	2/A	III	26	125	*B. paranthracis* MN5 (97.5)
FOOD_10_19_16_RSNT2_2A_R9-6399	Food	Restaurant 2	19-Oct	2/A	III	26	125	*B. paranthracis* MN5 (97.6)
FOOD_10_19_16_RSNT3_1E_R9-6407	Food	Restaurant 3	19-Oct	1/E	III	26	125	*B. paranthracis* MN5 (97.5)
FOOD_10_19_16_RSNT3_2A_R9-6403	Food	Restaurant 3	19-Oct	2/A	III	26	125	*B. paranthracis* MN5 (97.5)
FOOD_10_19_16_RSNT3_2A_R9-6404	Food	Restaurant 3	19-Oct	2/A	III	26	125	*B. paranthracis* MN5 (97.5)
FOOD_10_19_16_RSNT3_2A_R9-6405	Food	Restaurant 3	19-Oct	2/A	III	26	125	*B. paranthracis* MN5 (97.5)
FOOD_10_19_16_RSNT4_2B_R9-6408	Food	Restaurant 4	19-Oct	2/B	III	26	125	*B. paranthracis* MN5 (97.5)
FOOD_10_19_16_RSNT4_2B_R9-6409	Food	Restaurant 4	19-Oct	2/B	III	26	125	*B. paranthracis* MN5 (97.5)
FOOD_10_19_16_RSNT5_1C_R9-6411	Food	Restaurant 5	19-Oct	1/C	III	26	125	*B. paranthracis* MN5 (97.5)
HUMN_10_18_16_FECAL_NA_R9-6384	Human	Feces	18-Oct	NA	III	26	125	*B. paranthracis* MN5 (97.6)
HUMN_10_18_16_FECAL_NA_R9-6385	Human	Feces	18-Oct	NA	III	26	125	*B. paranthracis* MN5 (97.5)
HUMN_10_18_16_FECAL_NA_R9-6412	Human	Feces	18-Oct	NA	III	26	125	*B. paranthracis* MN5 (97.5)
HUMN_10_19_16_FECAL_NA_R9-6381	Human	Feces	19-Oct	NA	III	26	125	*B. paranthracis* MN5 (97.5)
HUMN_10_19_16_FECAL_NA_R9-6382	Human	Feces	19-Oct	NA	III	26	125	*B. paranthracis* MN5 (97.5)
HUMN_10_19_16_FECAL_NA_R9-6383	Human	Feces	19-Oct	NA	III	26	125	*B. paranthracis* MN5 (97.5)
FOOD_10_19_16_RSNT3_1E_R9-6406	Food	Restaurant 3	19-Oct	1/E	IV	24	92	*B. cereus* ATCC 14579 (98.9)
FOOD_10_19_16_RSNT5_1C_R9-6410	Food	Restaurant 5	19-Oct	1/C	IV	24	92	*B. cereus* ATCC 14579 (98.9)
HUMN_10_26_16_FECAL_NA_R9-6413	Human	Feces	26-Oct	NA	IV	142	92	*B. cereus* ATCC 14579 (98.8)

[a] Production date is designated by either 1 or 2; batch is one of A through H

[b] *panC* clade assigned *in silico* using BTyper 2.2.0

[c] Multi-locus sequence typing (MLST) sequence type (ST) assigned *in silico* using BTyper 2.2.0

[d] *rpoB* allelic type (AT) determined using Sanger sequencing and verified *in silico* using BTyper 2.2.0

[e] ANI, average nucleotide identity calculated using FastANI

which codes for emetic toxin-producing cereulide synthetase. In the case of isolate HUMN_10_18_16_FECAL_NA_R9-6384, *cesD* was split onto two contigs.

Based on average nucleotide identity (ANI) values, the three diarrheal group IV isolates were classified as *B. cereus s.s.* ($ANI > 95$; Table 5.3). The 30 emetic group III isolates from this outbreak, however, most closely resembled the type strain of *B. paranthracis* ($ANI > 95$; Table 5.3), indicating that the emetic group III and diarrheal group IV isolates from this outbreak are different *B. cereus* group species.

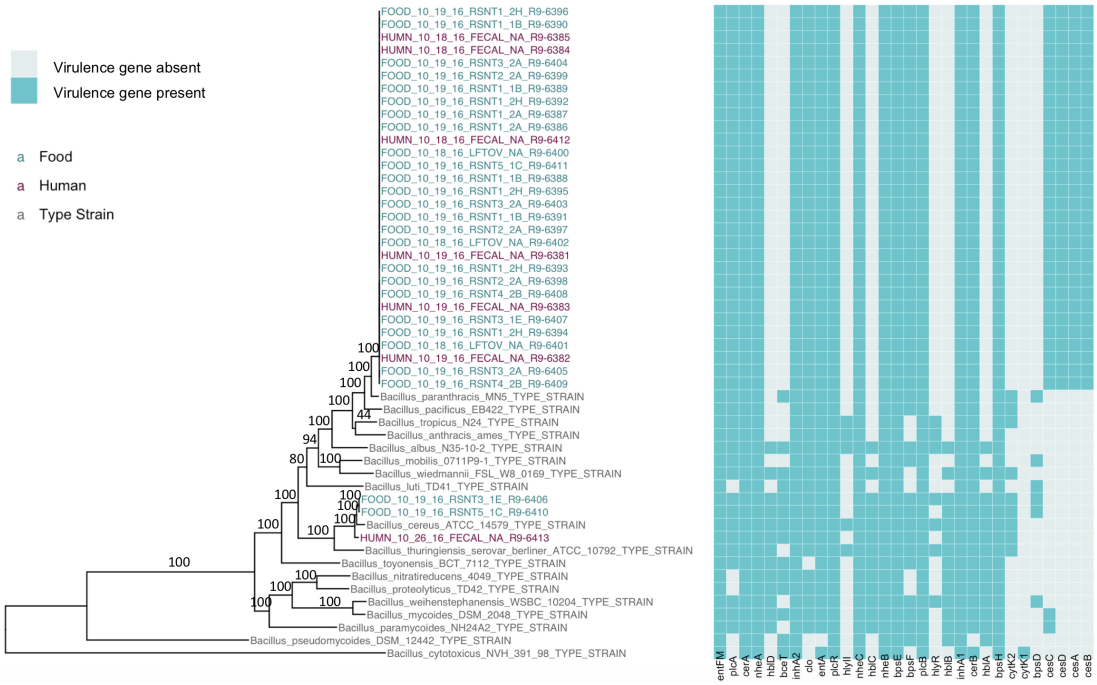


Figure 5.1: Maximum likelihood phylogeny of core SNPs identified in 33 isolates sequenced in conjunction with a *B. cereus* outbreak, as well as genomes of the 18 currently recognized *B. cereus* group species (shown in gray). Core SNPs were identified in all genomes using kSNP3. Heatmap corresponds to presence/absence of *B. cereus* group virulence genes detected in each sequence using BTyper. Tip labels in maroon and teal correspond to the seven human clinical isolates and 26 isolates from food sequenced in conjunction with this outbreak, respectively. Phylogeny is rooted at the midpoint, and branch labels correspond to bootstrap support percentages out of 500 replicates. Due to the short lengths and low bootstrap support (all values < 10) of branches within the outbreak clade, bootstrap support percentages are not shown on branches within the outbreak clade.

5.4.3 Emetic and Diarrheal *B. cereus* Isolates Associated With the Foodborne Outbreak do Not Differ in Cytotoxicity

All three diarrheal strains isolated in conjunction with the outbreak (FSL R9-6406, FSL R9-6410, and FSL R9-6413) were found to produce Hbl, as well as non-hemolytic enterotoxin (Nhe). Characterization of six representatives of the emetic isolates tested (i.e., FSL R9-6381, FSL R9-6382, FSL R9-6384, FSL R9-6389,

FSL R9-6395, and FSL R9-6399) revealed that they produced Nhe, but not Hbl. The supernatant of diarrheal *B. cereus s.s.* ATCC 14579 showed a stronger inhibitory effect on the viability of HeLa cells compared to supernatants of the 33 outbreak-associated isolates (Games-Howell $P < 0.05$; Figure 5.2). Furthermore, the viability of HeLa cells treated with 0.05% Triton X-100, the positive control, was significantly lower compared to viability of HeLa cells treated with bacterial supernatants (Games-Howell $P < 0.05$; Figure 5.2). Among all pairs of emetic isolates, only the viabilities of HeLa cells exposed to the supernatants of isolates FSL R9-6409 and FSL R9-6387 were found to differ (Games-Howell $P < 0.05$; Figure 5.2). The differences in HeLa cell viability after treatment with supernatants of these two emetic outbreak-associated strains are likely due to biological variability among replicates, as outbreak-associated emetic isolates were shown to be clonal (Figure 5.1). Taken together, the emetic group (represented by 30 emetic outbreak-associated isolates) had a mean cell viability of $97.5 \pm 5.1\%$, while the diarrheal group (represented by three diarrheal outbreak-associated isolates) gave a mean cell viability of $101.4 \pm 7.9\%$, as compared to the HeLa cells treated with BHI (i.e., negative control).

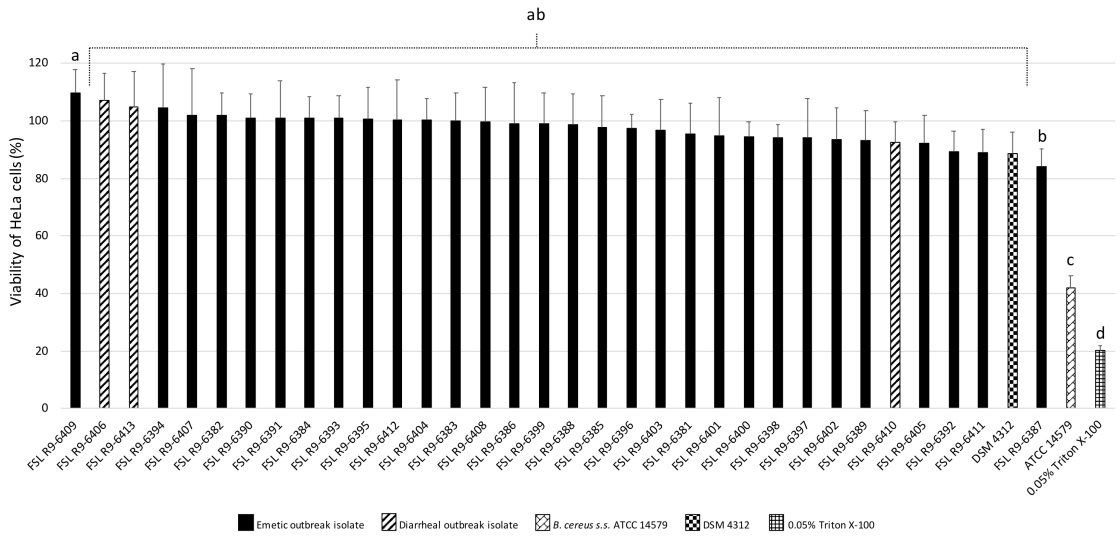


Figure 5.2: Percentage viability of HeLa cells when treated with supernatants of each isolate as determined by the WST-1 assay. Viability was calculated as ratio of corrected absorbance of solution when HeLa cells were treated with supernatants to the ratio of corrected absorbance of solution when HeLa cells were treated with BHI (i.e., negative control), converted to percentages. The columns represent the mean viabilities, while the error bars represent standard deviations for 12 technical replicates. Any two bars that do not share a common alphabetic character had significantly different percentage viability values ($P < 0.05$).

5.4.4 Core SNPs Identified Among *B. cereus* Group Outbreak Isolates From Two Phylogenetic Groups Are Dependent on Variant Calling Pipeline and Reference Genome Selection

To simulate a scenario in which genomes from a *B. cereus* outbreak spanning multiple phylogenetic groups were analyzed in aggregate, core SNPs were identified in all 33 outbreak isolates from groups III and IV (n = 30 and three isolates, respectively) using (i) combinations of five reference-based variant calling pipelines (Table 5.1) and three different reference genomes (Table 5.2) and

(ii) a reference-free SNP calling method (Table 5.1). When genomes from all 33 isolates were analyzed together, the number of core SNPs identified by each pipeline and reference combination varied by up to several orders of magnitude (Figure 5.3), often with little agreement between pipelines in terms of the core SNPs they reported (Figure 5.4). Independent of reference genome, the CFSAN pipeline was the most conservative, consistently identifying the fewest number of core SNPs when all 33 isolates were queried in aggregate (50, 27, and 0 core SNPs using reference genomes from groups III, IV, and VII, respectively) (Figure 5.3). This can be contrasted with the Samtools, Freebayes, and Parsnp pipelines, which produced upwards of 100,000 core SNPs when the selected reference genome was a member of one of the groups being queried in the outbreak isolate set (group III and IV; Figure 5.3). In cases where a distant genome was used as the reference (group VII *B. cytotoxicus* type strain chromosome), all reference-based pipelines reported fewer core SNPs than kSNP3's reference-free *k*-mer based SNP calling approach (Figure 5.3).

5.4.5 Choice of Variant Calling Pipeline Has Greater Influence on Core SNP Identification Than Choice of Closely Related Closed or Draft Reference Genome for Emetic Group III *B. cereus* Group Isolates

The 30 emetic group III isolates were queried in the absence of their group IV counterparts using combinations of five reference-based variant calling pipelines (Table 5.1) and two reference genomes (the closed chromosome of *B.*

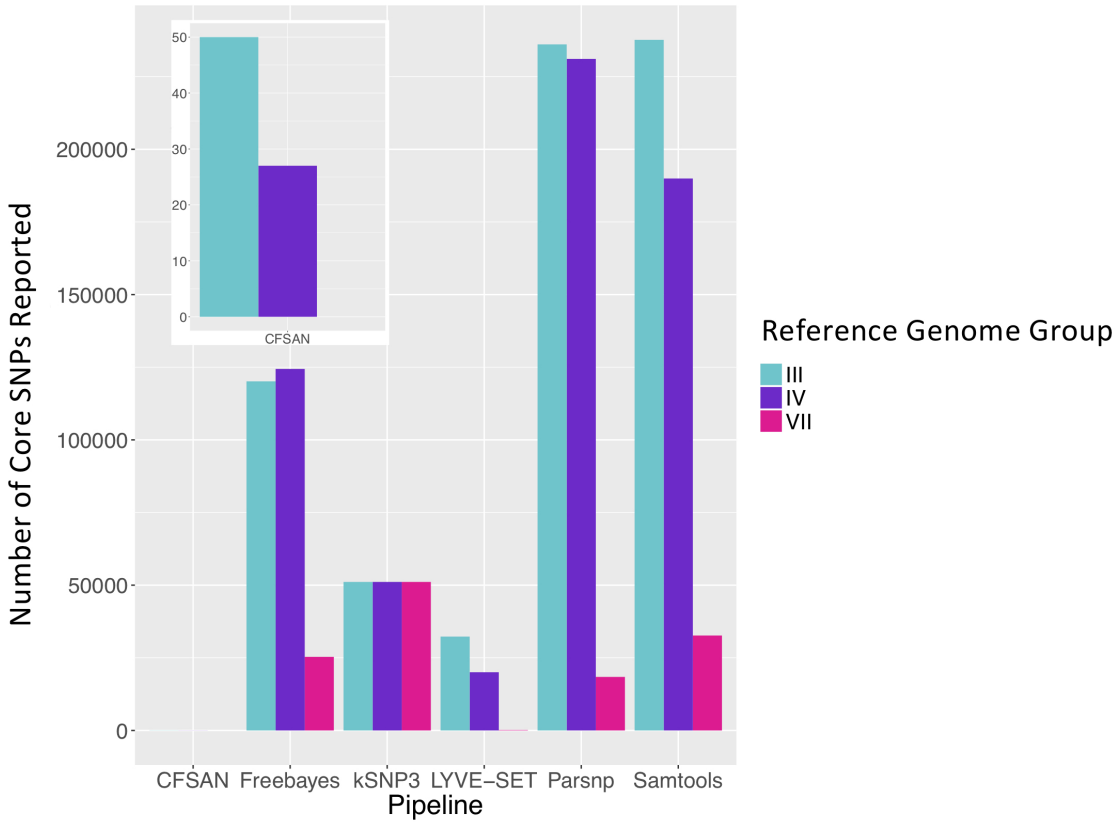


Figure 5.3: Number of core SNPs identified in 33 *B. cereus* group isolates from two phylogenetic groups (30 and 3 isolates from groups III and IV, respectively), sequenced in conjunction with a foodborne outbreak. Combinations of five reference-based variant calling pipelines and three reference genomes, as well as one reference-free SNP calling method (kSNP3), were tested.

cereus AH187, with and without dustmasking, and contigs of one of the isolates identified in this outbreak, with and without dustmasking; Table 5.2) and one reference-free SNP calling method (Table 5.1). In this scenario, the choice of variant calling pipeline had a greater effect on the number of core SNPs obtained than the choice of reference genome, as both reference genomes possessed the same virulence gene profile (virulotype), *rpoB* AT, *panC* group, MLST sequence type, and were of the same species (*B. paranthracis*; $ANI > 95$) as the 30 emetic isolates (Figure 5.5A). Congruent with this, the number of pairwise core SNP differences between emetic isolates sequenced in this outbreak varied

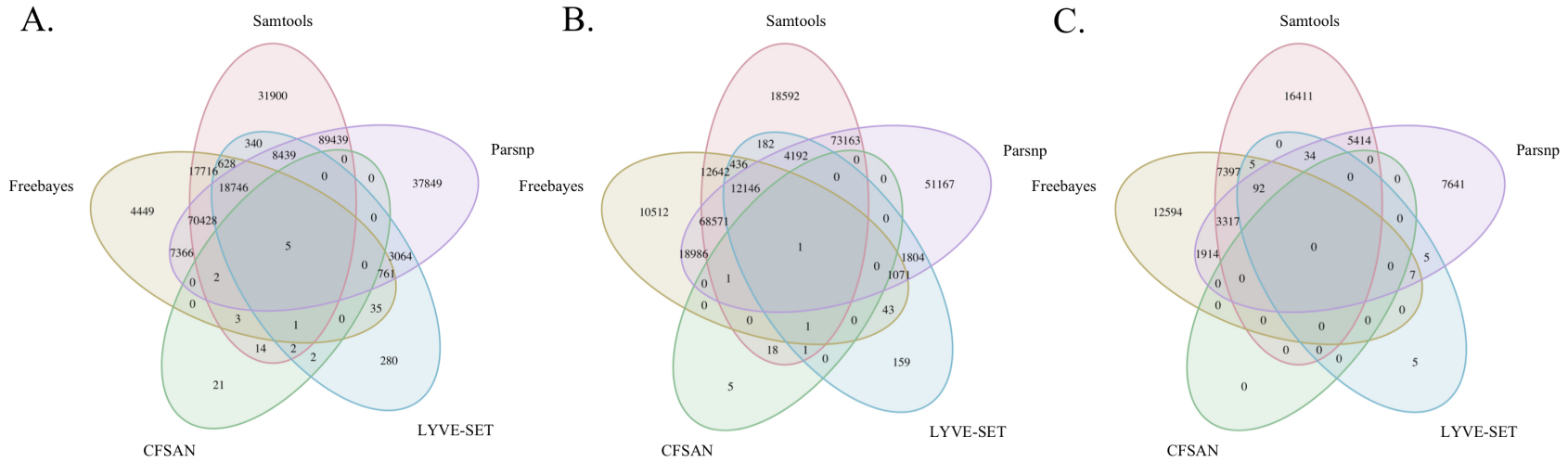


Figure 5.4: Comparison of core SNP positions reported by five reference-based variant-calling pipelines for 33 *B. cereus* group strains isolated in association with a foodborne outbreak, with the chromosomes of (A) *B. cereus* AH187 (group III), (B) *B. cereus s.s.* ATCC 14579 (group IV), and (C) *B. cytotoxicus* NVH 391-98 (group VII) used as reference genomes. Ellipses represent each pipeline.

more with the selection of variant calling pipeline than with reference genome (Figure 5.6A). When the unmasked closed chromosome of *B. cereus* AH187 was used as a reference, pairwise core SNP differences among emetic isolates from this outbreak ranged from 0 to 8 (mean of 2.9; CFSAN), 7 to 29 (mean of 16.1; Freebayes), 0 to 8 (mean of 2.8; LYVE-SET), 0 to 64 (mean of 23.6; Parsnp), and 1 to 16 SNPs (mean of 8.2; Samtools) (Figure 5.5A). Using the reference-free kSNP3 pipeline, this range was 1-46 SNPs (mean of 16.7; Figure 5.5A). The CFSAN and LYVE-SET pipelines produced nearly identical results in terms of the number and identity of the core SNPs called (23 and 22 SNPs, respectively, 20 of which were detected by both pipelines; Figure 5.7), as well as the topologies of the phylogenies those SNPs produced: all CFSAN and LYVE-SET phylogenies were more similar to each other than what would be expected by chance (Table 5.4 and Supplementary Table S4). Additionally, the two methods that relied on assembled genomes rather than short reads for SNP calling (kSNP3 and Parsnp) produced the greatest numbers of core SNPs (Figure 5.5A).

Within the emetic group III isolates associated with this outbreak, a total of 32 core SNPs were identified by two or more of the reference-based variant calling pipelines when the unmasked *B. cereus* AH187 genome was used as a reference, half of which were identified by all five pipelines (Figure 5.7). Out of these 32 SNPs, 23 were identified in protein coding genes, 14 of which produced non-synonymous amino acid changes (Supplementary Table S5). Genes with non-synonymous changes were involved in molybdopterin biosynthesis (WP_000544623.1), proteolysis (WP_000215096.1 and WP_000857793.1), chitin binding (WP_000795732.1), iron-hydroxamate transport (WP_000728195.1), DNA repair (WP_000947749.1 and WP_000867556.1), DNA replication (WP_000867556.1 and WP_000435993.1), protein transport and

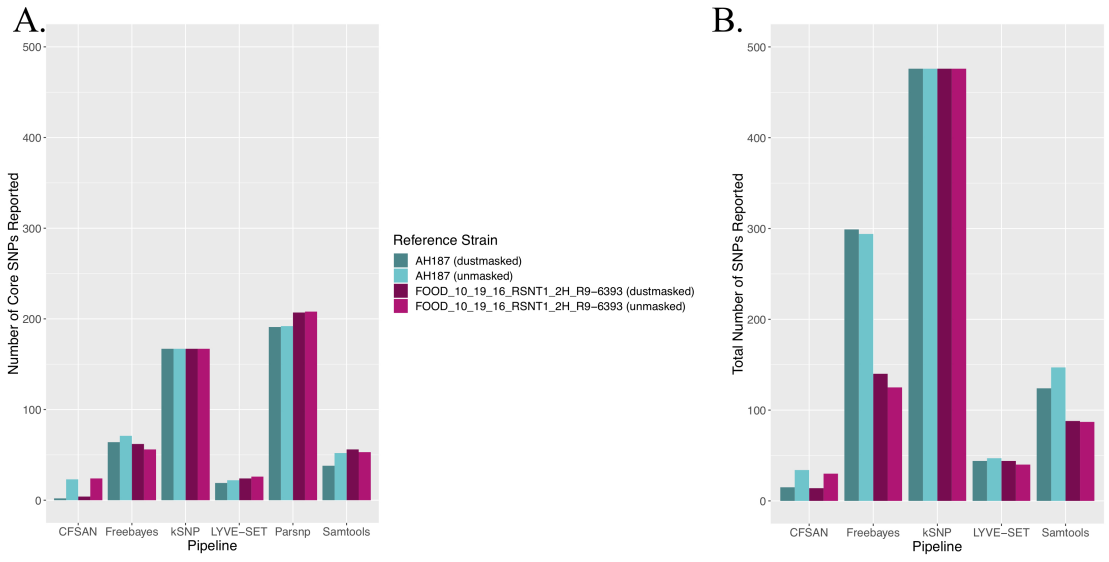


Figure 5.5: (A) Number of core SNPs and (B) total number of SNPs identified in 30 emetic *B. cereus* group III strains isolated in association with a foodborne outbreak. Combinations of (A) five and (B) four reference-based variant calling pipelines and two reference genomes (either dustmasked or unmasked) were tested, along with one reference-free SNP calling method (kSNP3). Because the Parsnp pipeline reports core SNPs by definition, it was excluded from Figure 5.5B (total SNPs). For quantification of the total number of SNPs (Figure 5.5B), all sites with more than one unique character were counted.

Table 5.4: Maximum likelihood phylogenies of 30 emetic group III outbreak isolates considered to be more topologically similar than would be expected by chance ($P < 0.05$).[a]

Reference Phylogeny[b]	*Query Phylogeny*[b]	*Corrected P-Value*[c]
AH187_CFSAN_NOdust_all	AH187_CFSAN_NOdust_core	0
AH187_CFSAN_NOdust_all	AH187_LYVE-SET_NOdust_all	0
AH187_CFSAN_NOdust_all	AH187_LYVE-SET_NOdust_core	0.0171
AH187_CFSAN_NOdust_all	AH187_LYVE-SET_YESdust_all	0
AH187_CFSAN_NOdust_all	AH187_LYVE-SET_YESdust_core	0.0171
AH187_CFSAN_NOdust_core	AH187_LYVE-SET_NOdust_all	0
AH187_CFSAN_NOdust_core	AH187_LYVE-SET_NOdust_core	0.0171
AH187_CFSAN_NOdust_core	AH187_LYVE-SET_YESdust_all	0
AH187_CFSAN_NOdust_core	AH187_LYVE-SET_YESdust_core	0.0171
AH187_Freebayes_NOdust_core	AH187_Freebayes_YESdust_core	0.0342
AH187_LYVE-SET_NOdust_all	AH187_LYVE-SET_NOdust_core	0.0171
AH187_LYVE-SET_NOdust_all	AH187_LYVE-SET_YESdust_all	0
AH187_LYVE-SET_NOdust_all	AH187_LYVE-SET_YESdust_core	0.0171
AH187_LYVE-SET_NOdust_core	AH187_LYVE-SET_YESdust_core	0
AH187_LYVE-SET_YESdust_all	AH187_LYVE-SET_YESdust_core	0.0171
AH187_Parsnp_NOdust_core	AH187_Parsnp_YESdust_core	0.0171

[a]Obtained from pairwise tests of tree topologies using a Z test based on the Kendall-Colijn metric; see Supplementary Table S4 for full table of comparisons

[b]Names of reference and query phylogenies denote reference genome ("AH187" for reference-based pipelines, "NOREF" for reference-free kSNP pipeline), pipeline ("CFSAN", "Freebayes", "kSNP", "LYVE-SET", "Parsnp", or "Samtools"), reference genome masking ("NOdust" for an unmasked reference genome, "YESdust" for a dustmasked reference genome, or "NAdust" for reference-free kSNP pipeline, for which dustmasking is not applicable), and SNPs used to construct the phylogeny ("core" for core SNPs, or "all" for core and accessory SNPs), separated by an underscore ("_")

[c]Bonferroni-corrected *P*-values for all tests that were significant at the $\alpha = 0.05$ level

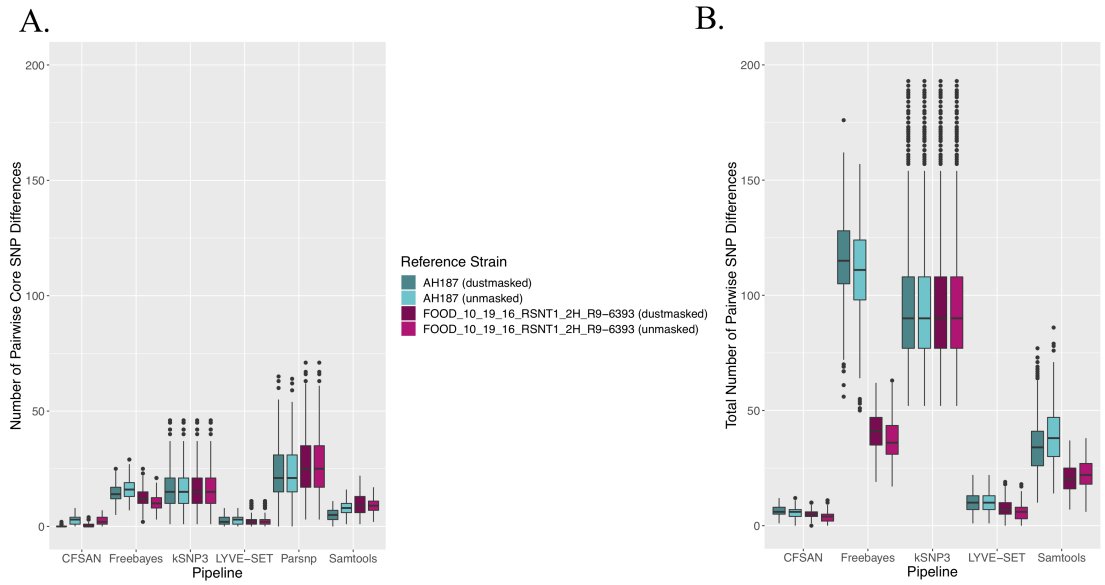


Figure 5.6: Ranges of pairwise (A) core SNP differences and (B) total SNP differences between 30 emetic group III *B. cereus* group strains isolated in conjunction with a foodborne outbreak. Combinations of (A) five and (B) four reference-based variant calling pipelines and two reference genomes (either dustmasked or unmasked), as well as one reference-free SNP calling method (kSNP3) were tested. Lower and upper box hinges correspond to the first and third quartiles, respectively. Lower and upper whiskers extend from the hinge to the smallest and largest values no more distant than 1.5 times the interquartile range from the hinge, respectively. Points represent pairwise distances that fall beyond the ends of the whiskers. Because the Parsnp pipeline reports core SNPs by definition, it was excluded from Figure 5.6B (pairwise differences in total SNPs). For quantification of pairwise differences in the total number of SNPs (Figure 5.6B), all sites with more than one unique character were included.

insertion into the membrane (WP_000727745.1), and glyoxylase/bleomycin resistance (WP_000800664.1).

In addition to detecting core SNPs in the genomes of the 30 emetic group III isolates, total (core and accessory) SNPs were detected in the 30 emetic group III genomes using combinations of four reference-based variant calling pipelines (Parsnp, which only reports core SNPs, was excluded; Table 5.1) and two reference genomes (the closed chromosome of *B. cereus* AH187 and contigs of one of the isolates identified in this outbreak, with and without dustmasking; Ta-

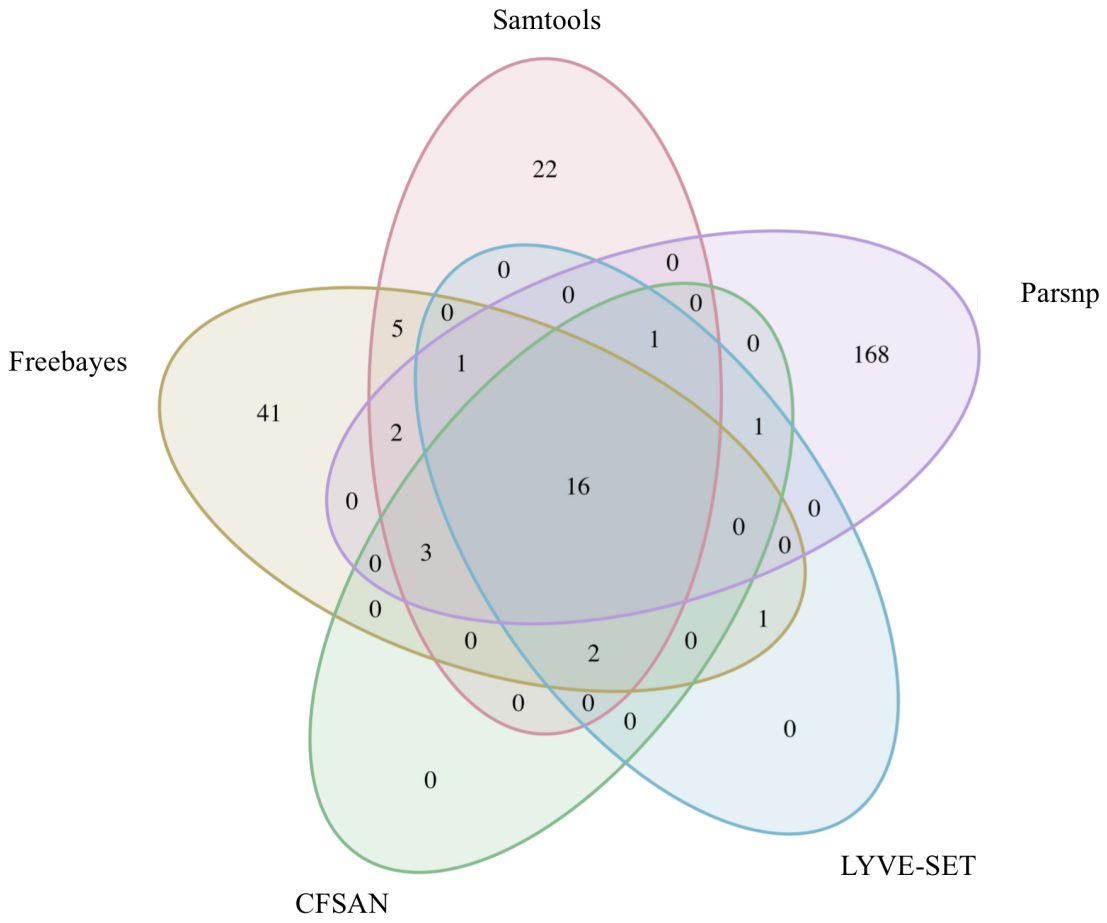


Figure 5.7: Comparison of core SNP positions reported by five variant-calling pipelines for 30 emetic group III *B. cereus* group outbreak isolates. Ellipses represent each pipeline, all of which used the chromosome of emetic group III *B. cereus* AH187 as a reference for variant calling.

ble 5.2) and one reference-free SNP calling method (Table 5.1). When total SNPs were accounted for, rather than solely core SNPs, all pipeline/reference genome combinations showed increases in the number of SNPs detected and the range of pairwise SNP differences between genomes (Figures 5.5B, 5.6B). Whether the addition of accessory SNPs translated into a significant difference in phylogenetic topology, however, depended on the variant calling pipeline used. When the *B. cereus* AH187 closed chromosome was used as a reference, SNPs detected using the LYVE-SET pipeline produced phylogenies considered to be

more topologically similar than would be expected by chance (Kendall-Colijn test $P < 0.05$), regardless of whether core SNPs or total SNPs were used to construct the phylogeny, and regardless of whether the *B. cereus* AH187 reference genome was dustmasked or not (Table 5.4 and Supplementary Table S4). Additionally, all phylogenies produced using the LYVE-SET pipeline and the *B. cereus* AH187 reference genome (i.e., each combination of core SNPs, total SNPs, dustmasked reference, and unmasked reference) were topologically similar to those produced using the CFSAN pipeline and the unmasked *B. cereus* AH187 reference genome, regardless of whether all SNPs were included or solely core SNPs (Table 5.4 and Supplementary Table S4). Other topologically similar phylogeny pairs included phylogenies constructed using (i) core SNPs identified with Freebayes, regardless of whether a dustmasked reference genome was used or not, and (ii) core SNPs identified with Parsnp, regardless of whether a dustmasked reference was used or not (Kendall-Colijn test $P < 0.05$; Table 5.4 and Supplementary Table S4).

5.4.6 Phylogenies Constructed Using Core SNPs Identified in 55 Emetic ST 26 *B. cereus* Genomes by kSNP3 and Parsnp Yield Similar Topologies

To compare the 30 emetic strains from this outbreak to other emetic group III isolates, all emetic group III assembled genomes with ST 26 were downloaded from NCBI. This produced a total of 55 emetic group III isolates with ST 26 (30 isolates from this outbreak and 25 from NCBI RefSeq). Among the 55 emetic ST 26 genomes, Parsnp identified almost twice as many core SNPs as kSNP3 (4,597

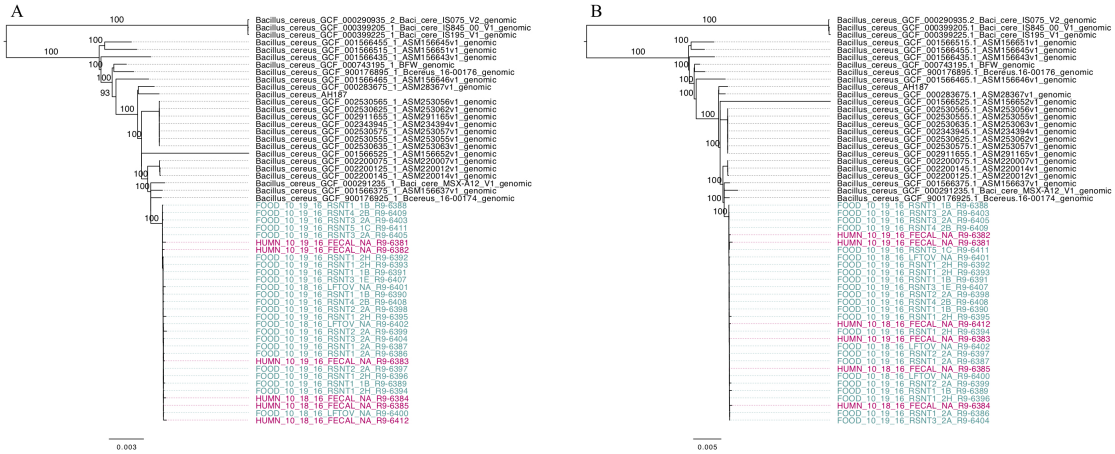


Figure 5.8: Maximum likelihood phylogenies of 30 emetic group III isolates (ST 26) sequenced in conjunction with a *B. cereus* outbreak, as well as all other emetic group III ST 26 genomes available in NCBI (n = 25; shown in black). Trees were constructed using core SNPs identified using (A) kSNP3 or (B) Parsnp. Tip labels in maroon and teal correspond to the six human clinical isolates and 24 isolates from food sequenced in conjunction with this outbreak, respectively. Branch labels correspond to bootstrap support percentages out of 1,000 replicates. Due to the short lengths and low bootstrap support of branches within the outbreak clade, bootstrap support percentages are not shown on branches within the outbreak clade.

and 2,593 core SNPs, respectively). However, the topologies of phylogenies produced using the core SNPs identified by each pipeline were found to be more similar than would be expected by chance (Kendall-Colijn test $P < 0.05$; Figure 5.8).

Based on pairwise core SNP differences, the publicly available genomes showed greater variability than the outbreak isolates described here, regardless of whether kSNP3 or Parsnp was used for variant calling (ANOVA-like permutation test $P < 0.05$; Supplementary Figure S1). Pairwise core SNP differences of the 30 emetic group III isolates from this outbreak ranged from 0 to 25 SNPs (mean of 8.3) and 0 to 44 SNPs (mean of 11.9) when the kSNP3 and Parsnp pipelines were used, respectively (Supplementary Figure S1). For external ST 26 isolates not associated with this outbreak, pairwise core SNP differ-

ences ranged from 0 to 1,474 SNPs (mean of 425.7) and 0 to 3,111 SNPs (mean of 828.3) when kSNP3 and Parsnp were used, respectively (Supplementary Figure S1). Between these two groups (the 30 emetic isolates from this outbreak and the 25 external emetic ST 26 isolates), pairwise core SNP differences ranged from 73 to 1,258 SNPs (mean of 301.7; kSNP3) and 74 to 2,709 SNPs (mean of 528.0; Parsnp) (Supplementary Figure S1). Reflecting this, the average of the ranks of pairwise SNP distances within emetic isolates from this outbreak was less than the average of the ranks of pairwise SNP distances between the emetic isolates from this outbreak and the external ST 26 isolates (ANOSIM $P < 0.05$). This is likely a result of the differences in variance between the outbreak and external ST 26 isolates, as supported by the results of the ANOVA-like permutation test (Anderson and D. C. I. Walsh 2013).

5.5 Discussion

While *B. cereus* causes a considerable number of foodborne illness cases annually, outbreaks are rarely investigated with the methodological rigor (e.g., use of WGS) that is increasingly used for surveillance and outbreak investigations targeting other foodborne pathogens. A specific challenge in the U.S. is that, unlike for some other diseases, disease cases caused by *B. cereus* are typically not reportable, even though foodborne illnesses, regardless of etiology, are reportable in some states, including NY. This, combined with the typically mild course of *B. cereus* infection, means that human *B. cereus* isolates are rarely available for WGS. Furthermore, even if clinical *B. cereus* group isolates are available, WGS may not be used for isolate characterization in cases where infections are mild. Due to the availability of *B. cereus* isolates for seven human cases, the out-

break reported here presented a unique opportunity to pilot the use of WGS for investigation of *B. cereus* outbreaks. The data and approaches presented here will not only facilitate future investigation of other *B. cereus* outbreaks but will also help with application of WGS for investigation of other foodborne disease outbreaks where limited reference WGS data and information on genomic diversity are available.

5.5.1 Addressing the Microbiological and Epidemiological Challenges Associated With Determining the Causative Agent of an Emetic Foodborne Outbreak

The agar MYP used for isolation of strains from food and human clinical samples in the outbreak reported here is one of the two selective differential agars recommended in the FDA BAM protocol for the isolation of *B. cereus* group strains (Tallent, Rhodehamel, et al. 1998). The second recommended agar, Bacara, has been shown to be more selective and more effective in suppressing the growth of other Gram-positive microorganisms that may be present in tested samples (e.g., other *Bacillus* species, *Listeria*, *Staphylococcus*) (Tallent, Kotewicz, et al. 2012; Kabir et al. 2017). Since Bacara medium has a proprietary formula and cannot be purchased in a dehydrated powder form (Tallent, Rhodehamel, et al. 1998), it is less likely to be readily available for use in labs that do not routinely test for *B. cereus* group species. Use of both types of media may increase the success of *B. cereus* group isolation from food and clinical samples, especially isolation of emetic strains (Ehling-Schulz, Svensson, et al. 2005; Ceuppens, Boon, and Uyttendaele 2013). Furthermore, the isolation of *B. cereus*

group strains associated with this outbreak was carried out at 37°C, which is higher than the temperature of 30°C that is recommended by the FDA BAM (Tallent, Rhodehamel, et al. 1998). Nevertheless, while incubation at this temperature may inhibit the growth of psychrotolerant species of the *B. cereus* group (e.g., *B. weihenstephanensis*), it is not expected to interfere with the isolation of *B. cereus* group strains that are able to grow at human body temperature and cause toxicoinfection. It is also not expected to compromise isolation of emetic isolates with the capacity to cause intoxication, as emetic strains have been previously found primarily in phylogenetic group III, which does not contain psychrotolerant strains (Carroll et al. 2017). Overall, use of both types of isolation media and a moderate incubation temperature of 30°C may minimize the isolation bias.

While the isolation of *B. cereus* group strains from food and clinical samples is essential for linking them to a potential foodborne outbreak, further information is needed to definitively prove that an outbreak was caused by *B. cereus*. Emetic disease caused by members of the *B. cereus* group can be attributed to the production of the highly heat- and pH-resistant toxin cereulide in food prior to ingestion (Ehling-Schulz, Fricker, and Scherer 2004; Ehling-Schulz, Frenzel, and Gohar 2015; Stenfors Arnesen, Fagerlund, and Granum 2008). Because cereulide is produced within the food matrix itself, prior to consumption, the mere presence of emetic *B. cereus* group strains in food or human clinical samples cannot definitively prove that an outbreak was caused by a member of the *B. cereus* group; rather, the presence of cereulide itself is essential for linking food and clinical samples to an outbreak with high confidence (Andersson et al. 2004; Stenfors Arnesen, Fagerlund, and Granum 2008). For this outbreak, the presence of cereulide in food and human clinical samples linked to the outbreak was not assessed, as testing for cereulide is not currently included

in the BAM protocol as a routine method for the detection and enumeration of *B. cereus* in food. Ergo, there is no definitive proof that the outbreak was caused by cereulide-producing emetic group III *B. cereus* and not a similar foodborne pathogen (e.g., enterotoxins produced by *Staphylococcus aureus*, which manifest in similar symptoms to those associated with cereulide) (Messelhausser et al. 2014). However, due to the presence of highly clonal, *ces*-positive group III ST 26 *B. cereus* group isolates among food and clinical samples linked to the outbreak, as well as epidemiological data that support this, the emetic strain is the most probable causative agent. While it is not currently included in the BAM protocol for *B. cereus* isolation (Tallent, Rhodehamel, et al. 1998), testing for the presence of cereulide in food and clinical samples linked to potential outbreaks caused by emetic *B. cereus* can aid in providing a definitive link between illness and causative agent.

5.5.2 Considerations for Addressing the Unique Challenges Associated With Characterization of Foodborne Outbreaks Linked to the *B. cereus* Group Using WGS

In *B. cereus* outbreaks, interpretation of WGS data can be challenging, especially in cases where strains of multiple closely related species or subtypes appear to be associated with an outbreak. *B. cereus* outbreaks, particularly emetic outbreaks caused by cereulide-producing *B. cereus* group isolates, are often associated with improper handling of food (e.g., temperature abuse) (Ehling-Schulz, Fricker, and Scherer 2004; Stenfors Arnesen, Fagerlund, and Granum 2008). This, and their ubiquitous presence in the environment, make it important to

consider the possibility of a multi-strain or multi-species outbreak in addition to a single-source outbreak caused by a single strain. In the outbreak characterized here, *B. cereus* group strains from two phylogenetic groups, III and IV, were isolated from both human clinical stool samples, as well as refried beans linked to the outbreak. The separation of outbreak-related isolates into three diarrheal group IV isolates (representing two distinct STs) and 30 emetic isolates may be explained by one of the following scenarios: (i) the outbreak was caused by refried beans contaminated with multiple *B. cereus* group species (isolates from groups III and IV), both of which caused illness in humans, (ii) in addition to housing emetic outbreak strains that belonged to group III, samples of refried beans and patient stool samples harbored group IV *B. cereus s.l.* isolates that were not part of the outbreak but were incidentally isolated from stool and food samples, or (iii) a subset of patient stool samples and food samples did not harbor *B. cereus s.l.* group III isolates belonging to the outbreak, but did harbor group IV strains that were isolated and sequenced. In order to determine which of these scenarios explains the presence of multiple *B. cereus* group species among isolates sequenced in conjunction with a foodborne outbreak, additional epidemiological and microbiological data are needed.

Valuable metrics for inclusion/exclusion of *B. cereus* group cases in a foodborne outbreak include patient exposure, patient symptoms (e.g., vomiting, diarrhea, onset and duration of illness), levels of *B. cereus* present in implicated food and patient samples (CFU/g or CFU/ml), cytotoxicity of isolates, and the approach used to select bacterial colonies to undergo WGS (Glasset et al. 2016). However, some of these data may be more valuable than others. In their characterization of 564 *B. cereus* group strains associated with 140 "strong-evidence" foodborne outbreaks in France between 2007 and 2014, Glasset et al. (Glasset et

al. 2016) found that patient symptoms could not be associated with the presence of emetic and diarrheal strains. More than half (57%) of the *B. cereus* outbreaks queried in their study included patients exhibiting both emetic and diarrheal symptoms. Similar results were observed here, as emetic and diarrheal symptoms were reported in 88 and 38% of cases, respectively, with both vomiting and diarrhea reported by multiple patients. All emetic isolates associated with this outbreak carried *nhe* genes and also produced Nhe enterotoxin, as determined using the immunoassay. While it has been proposed that a combination of emetic and diarrheal symptoms may be due to the fact that emetic group III isolates have been shown to produce diarrheal enterotoxin Nhe at high levels (Glasset et al. 2016), incongruences between isolate virulotype and patient symptoms may still exist. Importantly, this indicates the need for further investigation of factors affecting the expression of *B. cereus* group virulence genes, as well as their potential synergistic activities (Doll, Ehling-Schulz, and Vogelmann 2013).

Another metric that can be used for determining whether *B. cereus* group isolates are part of an outbreak or not is the level of *B. cereus* present in the implicated food. Like patient symptoms, *B. cereus* counts from implicated foods may aid in an outbreak investigation, but likely cannot definitively prove whether an isolate is part of an outbreak or not. For example, outbreaks caused by implicated foods with *B. cereus* counts of $< 10^3$ CFU/g and as low as 400 CFU/g for diarrheal and emetic diseases, respectively, have been described (Glasset et al. 2016), despite levels of at least 10^5 CFU/g often being detected in implicated foods (Stenfors Arnesen, Fagerlund, and Granum 2008). The levels of *B. cereus* present in refried beans in the outbreak described here were not determined. However, like patient symptoms, *B. cereus* count data may be a useful

supplemental metric for investigating *B. cereus* group outbreaks in the future.

In addition to patient symptoms and pathogen load in the food, incubation period can be used to determine whether an isolate is part of an outbreak or not, as it is significantly shorter for emetic strains than diarrheal strains (Ehling-Schulz, Fricker, and Scherer 2004; Stenfors Arnesen, Fagerlund, and Granum 2008; Glasset et al. 2016). In the outbreak described here, the patient from which a non-emetic group IV *B. cereus* group strain was isolated reported an incubation time of 1 h, the lowest incubation time of all seven confirmed human clinical cases. However, this is still within the observed range of incubation times for emetic *B. cereus* disease (0.5-6 h) (Stenfors Arnesen, Fagerlund, and Granum 2008). Although no emetic group III *B. cereus s.l.* strain was isolated from the clinical sample, it is possible that the patient could have been intoxicated with cereulide produced in the food by the emetic *B. cereus* strain that caused the outbreak. However, it is also possible that a pathogen which causes similar symptoms to foodborne illness caused by emetic *B. cereus* was responsible for the patient's illness (e.g., *Staphylococcus aureus*).

Lastly, cytotoxicity data may also be leveraged to include/exclude outbreak-associated *B. cereus* group isolates. In the outbreak described here, the patient from which a non-emetic group IV *B. cereus* group strain was isolated reported vomiting and nausea and no diarrheal symptoms, despite the clinical isolate's possession of multiple diarrheal toxin genes and no emetic toxin genes. This could suggest that the patient was intoxicated with the cereulide, but the isolate itself did not survive the passage through the patient's gastrointestinal tract, or that it survived in a low concentration that resulted in failure of isolation on MYP. It is also possible that our understanding of the specific virulence genes re-

sponsible for different *B. cereus*-associated disease symptoms is still incomplete and that the diarrheal isolate obtained from the clinical sample was in fact responsible for symptoms of vomiting and nausea. To further investigate this, we carried out immunoassay-based detection of Hbl and Nhe enterotoxins, as well as a WST-1 proliferation assay with HeLa cells exposed to bacterial supernatants presumably containing toxins. The results of Hbl and Nhe immunodetection and cytotoxicity revealed that diarrheal isolates only had mild detrimental effects on HeLa cell viability, despite the fact that they produced both hemolysin BL and non-hemolytic enterotoxins. This can be contrasted with the *B. cereus s.s.* type strain, which substantially reduced the viability of the HeLa cells.

For the outbreak described here, results obtained using a combination of microbiological, epidemiological, and bioinformatic methods indicate that hypothesis (i), in which the diarrheal strains were part of a multi-species outbreak, can likely be excluded. Evidence supporting the conclusion that the human clinical diarrheal isolate was not part of the outbreak described here include: (i) the emetic symptoms reported by the patient were incongruent with the virulotype of the isolate, (ii) the incubation time was typical for intoxication, (iii) the human clinical diarrheal isolate had a different MLST ST compared to all other isolates sequenced in this outbreak, and (iv) the human diarrheal isolate did not exhibit substantial cytotoxicity against HeLa cells (Figure 5.2). This may be due to the fact that this case was not part of the outbreak and was due to an infection or intoxication caused by another pathogen that leads to disease symptoms similar to *B. cereus* (e.g., *Staphylococcus aureus*), or that a group IV *B. cereus* strain was isolated and sequenced in lieu of the group III emetic outbreak isolate. There is limited evidence as to whether humans can be asymptomatic carriers of group IV *B. cereus* (Ghosh 1978; Turnbull and Kramer 1985), making

it likely that isolation and sequencing of a group IV *B. cereus* strain could be due to the use of MYP agar as the sole selective agar, which has been shown to hinder detection of emetic *B. cereus* group isolates (Ehling-Schulz, Svensson, et al. 2005; Ceuppens, Boon, and Uyttendaele 2013). In future outbreaks, the use of additional selective media (e.g., Bacara agar), enrichment media, and isolation temperatures may aid in isolation of the causative *B. cereus* group strain.

While we have shown here that WGS data can be a valuable tool for characterizing *B. cereus* group isolates from a foodborne outbreak, our results also showcase the importance of supplementing WGS data with epidemiological and microbiological metadata to draw meaningful conclusions from *B. cereus* group genomic data. Furthermore, the availability of WGS and cytotoxicity data from a larger set of *B. cereus* isolates from symptomatic patients may also provide an opportunity to use comparative genomics approaches to further explore virulence genes that are linked to different disease outcomes in the future.

5.5.3 Recommendations for Analyzing Illumina WGS Data From *B. cereus* Group Isolates Potentially Linked to a Foodborne Outbreak

WGS is being used increasingly to characterize isolates associated with foodborne disease cases and outbreaks, and rightfully so, as it offers the ability to characterize foodborne pathogens at unprecedented resolution, and it has been able to improve outbreak and cluster detection for numerous foodborne pathogens (Allard et al. 2017; Jasna Kovac et al. 2017; Moran-Gilad 2017;

Taboada et al. 2017), including *Salmonella enterica* (Taylor et al. 2015; Hoffmann et al. 2016; Gymoese et al. 2017), *Escherichia coli* (Grad et al. 2012; Holmes et al. 2015; Rusconi et al. 2016), and *Listeria monocytogenes* (Jackson et al. 2016; Kwong et al. 2016; Chen, Luo, Pettengill, et al. 2017; Chen, Luo, Curry, et al. 2017; Moura et al. 2017). However, as demonstrated here and elsewhere, variant calling pipelines and the various mapping/alignment, SNP calling, and SNP filtering practices that they employ (e.g., removal of recombination and clustered SNPs) can influence the identification of SNPs in WGS data and, thus, the topology of a resulting phylogeny (Pightling, Petronella, and Pagotto 2014; Pightling, Petronella, and Pagotto 2015; Croucher et al. 2015; Hwang et al. 2015; Katz et al. 2017; Sandmann et al. 2017). This can be particularly problematic for outbreak and cluster detection in bacterial pathogen surveillance: pairwise SNP thresholds are currently widely used to make initial decisions regarding the inclusion or exclusion of isolates in a given outbreak (Taylor et al. 2015; Gymoese et al. 2017; Mair-Jenkins et al. 2017; McCloskey and Poon 2017; Walker et al. 2018). In such scenarios, just a few SNPs can be the deciding factor in whether a bacterial pathogen is included or excluded as part of an outbreak or cluster (Katz et al. 2017), rendering the choice of variant calling method as non-trivial. Furthermore, choosing an appropriate variant calling pipeline can be particularly challenging for pathogens where there are limited data and expertise with WGS, as is currently the case with *B. cereus*.

As demonstrated here, the choice of variant calling pipeline can greatly influence the number of core SNPs identified in *B. cereus* group isolates associated with a foodborne outbreak. In the case of a multi-group outbreak, this effect can be magnified. Naively calling variants in isolates that span multiple *B. cereus s.l.* phylogenetic groups in aggregate can lead to orders of magni-

tudes of difference in the number of core SNPs identified by different variant calling pipelines/reference genome combinations. In a multi-group outbreak scenario, it is essential to note that one is effectively dealing with genomic data from *multiple species* (i.e., $ANI < 95$), making it impossible to find a reference genome that is closely related to all isolates in a putative outbreak. In the case of some reference-based pipelines that are specifically tailored to identify variants in bacterial isolates from outbreaks (e.g., CFSAN, which is not suited for bacteria differing by more than a few hundred SNPs), calling variants in multiple groups or within a distant reference genome is inappropriate (Davis et al. 2015). Thus, querying outbreak isolates from multiple groups in aggregate using reference-based variant calling methods should be avoided. Furthermore, the results presented here showcase the value of employing single- and/or multi-locus typing approaches prior to variant calling, either via Sanger sequencing or *in silico* using tools, such as BTyper, as they can aid the design of downstream bioinformatics analyses, including reference genome selection and data partitioning by phylogenetic group.

When the three phylogenetic group IV isolates were excluded from analyses, leaving only the emetic group III isolates, the selection of reference genome caused fewer core SNP discrepancies than choice of variant calling pipeline, provided the reference genome was "similar" to the genomes analyzed. While the selection of a reference genome for reference-based variant calling is not trivial (Pightling, Petronella, and Pagotto 2014; Olson et al. 2015), reference-based variant calling using a closed chromosome (*B. cereus* AH187) and a draft genome (FOOD_10_19_16_RSNT1_2H_R9-6393) from two isolates that were closely related to, or among the emetic group III isolates sequenced in this outbreak produced nearly identical results in terms of the number and identity of core

SNPs detected. Both reference genomes were identical to the emetic group III outbreak isolates sequenced here in terms of *panC* group, *rpoB* AT, MLST ST, and virulotype. Additionally, the closed chromosome and draft genome had ANI values of > 99.8 and 99.9, respectively, relative to all emetic group III outbreak isolates in this study, which can be considered highly similar. Comparable findings have been observed in analyses of *Salmonella enterica* serovar Heidelberg WGS data (Usongo et al. 2018), suggesting that either closed genomes or high-quality draft genomes are adequate for reference-based SNP calling, provided both are similar enough to the outbreak strains being queried. While the thresholds at which reference genomes become "similar enough" and of sufficient quality for reference-based SNP calling for outbreak detection warrant further investigation, we have demonstrated here that, for emetic group III ST 26 *B. cereus* group genomes, the publicly available closed chromosome of *B. cereus* AH187 can serve as an adequate standard.

With regard to differences in the number of core SNPs identified in the 30 emetic group III isolates using different variant calling pipelines, the pipelines that used assembled genomes as input (kSNP3 and Parsnp) produced higher numbers of core SNPs than their counterparts that relied on short Illumina reads. Additionally, when used to query core SNPs in 55 emetic group III ST 26 *B. cereus* group genomes, both kSNP3 and Parsnp produced core SNPs that yielded topologically similar phylogenies. kSNP3 employs a reference-free *k*-mer based SNP calling approach (Gardner and Hall 2013; Gardner, Slezak, and Hall 2015), while Parsnp uses a reference-based core genome alignment approach (Treangen et al. 2014), and both are useful for calling variants in large data sets. These approaches are also valuable when reads are not available for SNP calling (Olson et al. 2015), as demonstrated here by the comparison of

outbreak genomes with publicly available genomes: core SNPs obtained using both kSNP3 and Parsnp were able to consistently produce phylogenies in which the 30 emetic isolates from this outbreak formed a well-supported clade among all emetic group III ST 26 *B. cereus* group genomes. However, kSNP3 has been shown to lack specificity relative to other pipelines (i.e., CFSAN, LYVE-SET) when differentiating outbreak isolates from non-outbreak isolates for *L. monocytogenes*, *E. coli*, and *S. enterica* (Katz et al. 2017). Here, the CFSAN and LYVE-SET pipelines identified similar SNPs that produced highly congruent phylogenies. This is unsurprising, considering that both the CFSAN and LYVE-SET pipelines were designed specifically for identifying SNPs in closely related strains from outbreaks (Katz et al. 2017), and both employ the most stringent filtering criteria of all pipelines tested here.

5.5.4 As WGS Becomes Routinely Integrated Into Food Safety, Clinical, and Epidemiological Realms, It Is Likely That the Number of Illnesses Attributed to *B. cereus* Will Increase

Here, we offer the first description of a foodborne outbreak caused by *B. cereus* group species to be characterized using WGS, and we provide a glimpse into the genomic variation one might expect within an emetic group III *B. cereus* outbreak using several different variant calling pipelines. However, our ability to query emetic group III genomes outside of this outbreak is limited by the lack of publicly available genomic data and metadata from emetic isolates. Of the 2,156

B. cereus group genomes available in NCBI's RefSeq database as of March 2018, only 29 were from group III and possessed the *cesABCD* operon, 25 of which belonged to MLST ST 26. While not ideal, this is an improvement, as there were only 19 emetic group III genomes available in NCBI's Genbank database in April 2017 (Carroll et al. 2017). As more *B. cereus* group WGS data, particularly, data from emetic *B. cereus* group isolates, become publicly available, more outbreaks and clusters are likely to be resolved in tandem, a phenomenon that has been observed for *L. monocytogenes* (Jackson et al. 2016). Additionally, variant calling and cluster/outbreak detection methods for characterizing *B. cereus* group isolates from foodborne outbreaks can be further refined and optimized as more WGS, metadata and epidemiological data become available for clinical and non-clinical isolates.

5.6 Acknowledgments

This material is based on work supported by the National Science Foundation Graduate Research Fellowship Program under grant no. DGE-1144153. This work was supported also by the USDA National Institute of Food and Agriculture Hatch Appropriations under Project #PEN04646 and Accession #1015787, and Penn State Huck Institutes of the Life Sciences that supported the whole-genome sequencing through the Penn State Genomics Core Facility. The authors would like to acknowledge the Wadsworth Center Tissue Culture & Media Core for providing the media used in this work, and Dr. Joshua Lambert from The Pennsylvania State University for providing tissue culture laboratory facility and advising.

5.7 References

Allard, M. W. et al. (2017). "Genomics of foodborne pathogens for microbial food safety". In: *Curr Opin Biotechnol* 49, pp. 224–229. DOI: 10.1016/j.copbio.2017.11.002.

Anderson, M. J. and D. C. I. Walsh (2013). "PERMANOVA, ANOSIM, and the Mantel test in the face of heterogeneous dispersions: What null hypothesis are you testing?" In: *Ecological Monographs* 83.4, pp. 557–574. DOI: 10.1890/12-2010.1.

Andersson, M. A. et al. (2004). "Sperm bioassay for rapid detection of cereulide-producing *Bacillus cereus* in food and related environments". In: *Int J Food Microbiol* 94.2, pp. 175–83. DOI: 10.1016/j.ijfoodmicro.2004.01.018.

Ashton, Philip et al. (2015). "Revolutionising Public Health Reference Microbiology using Whole Genome Sequencing: *Salmonella* as an exemplar". In: *bioRxiv*. DOI: 10.1101/033225.

Bankevich, A. et al. (2012). "SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing". In: *J Comput Biol* 19.5, pp. 455–77. DOI: 10.1089/cmb.2012.0021.

Bennett, S. D., K. A. Walsh, and L. H. Gould (2013). "Foodborne disease outbreaks caused by *Bacillus cereus*, *Clostridium perfringens*, and *Staphylococcus aureus*–United States, 1998-2008". In: *Clin Infect Dis* 57.3, pp. 425–33. DOI: 10.1093/cid/cit244.

Bolger, A. M., M. Lohse, and B. Usadel (2014). "Trimmomatic: a flexible trimmer for Illumina sequence data". In: *Bioinformatics* 30.15, pp. 2114–20. DOI: 10.1093/bioinformatics/btu170.

Bruen, T. C., H. Philippe, and D. Bryant (2006). "A simple and robust statistical test for detecting the presence of recombination". In: *Genetics* 172.4, pp. 2665–81. DOI: 10.1534/genetics.105.048975.

Carroll, L. M., J. Kovac, R. A. Miller, and M. Wiedmann (2017). "Rapid, high-throughput identification of anthrax-causing and emetic *Bacillus cereus* group genome assemblies using BTyper, a computational tool for virulence-based classification of *Bacillus cereus* group isolates using nucleotide sequencing data". In: *Appl Environ Microbiol*. DOI: 10.1128/AEM.01096-17.

Castiaux, V., X. Liu, L. Delbrassinne, and J. Mahillon (2015). "Is Cytotoxin K from *Bacillus cereus* a bona fide enterotoxin?" In: *Int J Food Microbiol* 211, pp. 79–85. DOI: 10.1016/j.ijfoodmicro.2015.06.020.

Ceuppens, S., N. Boon, and M. Uyttendaele (2013). "Diversity of *Bacillus cereus* group strains is reflected in their broad range of pathogenicity and diverse ecological lifestyles". In: *FEMS Microbiol Ecol* 84.3, pp. 433–50. DOI: 10.1111/1574-6941.12110.

Chen, Y., Y. Luo, P. Curry, et al. (2017). "Assessing the genome level diversity of *Listeria monocytogenes* from contaminated ice cream and environmental samples linked to a listeriosis outbreak in the United States". In: *PLoS One* 12.2, e0171389. DOI: 10.1371/journal.pone.0171389.

Chen, Y., Y. Luo, J. Pettengill, et al. (2017). "Singleton Sequence Type 382, an Emerging Clonal Group of *Listeria monocytogenes* Associated with Three Multistate Outbreaks Linked to Contaminated Stone Fruit, Caramel Apples, and Leafy Green Salad". In: *J Clin Microbiol* 55.3, pp. 931–941. DOI: 10.1128/JCM.02140-16.

Clarke, K. R. (1993). "Non-parametric multivariate analyses of changes in community structure". In: *Australian Journal of Ecology* 18.1, pp. 117–143. DOI: 10.1111/j.1442-9993.1993.tb00438.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1442-9993.1993.tb00438.x.

Croucher, N. J. et al. (2015). "Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins". In: *Nucleic Acids Res* 43.3, e15. DOI: 10.1093/nar/gku1196.

Danecek, P. et al. (2011). "The variant call format and VCFtools". In: *Bioinformatics* 27.15, pp. 2156–8. DOI: 10.1093/bioinformatics/btr330.

Davis, Steve et al. (2015). "CFSAN SNP Pipeline: an automated method for constructing SNP matrices from next-generation sequence data". In: *PeerJ Computer Science* 1, e20. DOI: 10.7717/peerj-cs.20.

de Jonge, Edwin (2018). *docopt: Command-Line Interface Specification Language*. R package version 0.6.1.

Doll, V. M., M. Ehling-Schulz, and R. Vogelmann (2013). "Concerted action of sphingomyelinase and non-hemolytic enterotoxin in pathogenic *Bacillus cereus*". In: *PLoS One* 8.4, e61404. DOI: 10.1371/journal.pone.0061404.

Ehling-Schulz, M., E. Frenzel, and M. Gohar (2015). "Food-bacteria interplay: pathometabolism of emetic *Bacillus cereus*". In: *Front Microbiol* 6, p. 704. DOI: 10.3389/fmicb.2015.00704.

Ehling-Schulz, M., M. Fricker, and S. Scherer (2004). "*Bacillus cereus*, the causative agent of an emetic type of food-borne illness". In: *Mol Nutr Food Res* 48.7, pp. 479–87. DOI: 10.1002/mnfr.200400055.

Ehling-Schulz, M., B. Svensson, et al. (2005). "Emetic toxin formation of *Bacillus cereus* is restricted to a single evolutionary lineage of closely related strains". In: *Microbiology* 151.Pt 1, pp. 183–97. DOI: 10.1099/mic.0.27607-0.

Fisichella, M. et al. (2009). "Mesoporous silica nanoparticles enhance MTT formazan exocytosis in HeLa cells and astrocytes". In: *Toxicol In Vitro* 23.4, pp. 697–703. DOI: 10.1016/j.tiv.2009.02.007.

Gardner, S. N. and B. G. Hall (2013). "When whole-genome alignments just won't work: kSNP v2 software for alignment-free SNP discovery and phylogenetics of hundreds of microbial genomes". In: *PLoS One* 8.12, e81760. DOI: 10.1371/journal.pone.0081760.

Gardner, S. N., T. Slezak, and B. G. Hall (2015). "kSNP3.0: SNP detection and phylogenetic analysis of genomes without genome alignment or reference genome". In: *Bioinformatics* 31.17, pp. 2877–8. DOI: 10.1093/bioinformatics/btv271.

Garrison, Erik and Gabor Marth (2012). "Haplotype-based variant detection from short-read sequencing". In: *arXiv* 1207.3907v2.

Ghosh, A. C. (1978). "Prevalence of *Bacillus cereus* in the faeces of healthy adults". In: *J Hyg (Lond)* 80.2, pp. 233–6.

Glasset, B. et al. (2016). "*Bacillus cereus*-induced food-borne outbreaks in France, 2007 to 2014: epidemiology and genetic characterisation". In: *Euro Surveill* 21.48. DOI: 10.2807/1560-7917.ES.2016.21.48.30413.

Grad, Y. H. et al. (2012). "Genomic epidemiology of the *Escherichia coli* O104:H4 outbreaks in Europe, 2011". In: *Proc Natl Acad Sci U S A* 109.8, pp. 3065–70. DOI: 10.1073/pnas.1121491109.

Granum, P. E. and T. Lund (1997). "*Bacillus cereus* and its food poisoning toxins". In: *FEMS Microbiol Lett* 157.2, pp. 223–8.

Guinebretiere, M. H., S. Auger, et al. (2013). "*Bacillus cytotoxicus* sp. nov. is a novel thermotolerant species of the *Bacillus cereus* Group occasionally associated with food poisoning". In: *Int J Syst Evol Microbiol* 63.Pt 1, pp. 31–40. DOI: 10.1099/ijs.0.030627-0.

Guinebretiere, M. H., F. L. Thompson, et al. (2008). "Ecological diversification in the *Bacillus cereus* Group". In: *Environ Microbiol* 10.4, pp. 851–65. DOI: 10.1111/j.1462-2920.2007.01495.x.

Guinebretiere, M. H., P. Velge, et al. (2010). "Ability of *Bacillus cereus* group strains to cause food poisoning varies according to phylogenetic affiliation (groups I to VII) rather than species affiliation". In: *J Clin Microbiol* 48.9, pp. 3388–91. DOI: 10.1128/JCM.00921-10.

Gymoese, P. et al. (2017). "Investigation of Outbreaks of *Salmonella enterica* Serovar Typhimurium and Its Monophasic Variants Using Whole-Genome Sequencing, Denmark". In: *Emerg Infect Dis* 23.10, pp. 1631–1639. DOI: 10.3201/eid2310.161248.

Heibl, C. (2008 onwards). *PHYLOCH: R language tree plotting tools and interfaces to diverse phylogenetic software packages. http://www.christophheibl.de/Rpackages.html*.

Hoffmann, M. et al. (2016). "Tracing Origins of the *Salmonella* Bareilly Strain Causing a Food-borne Outbreak in the United States". In: *J Infect Dis* 213.4, pp. 502–8. DOI: 10.1093/infdis/jiv297.

Holmes, A. et al. (2015). "Utility of Whole-Genome Sequencing of *Escherichia coli* O157 for Outbreak Detection and Epidemiological Surveillance". In: *J Clin Microbiol* 53.11, pp. 3565–73. DOI: 10.1128/JCM.01066-15.

Hwang, S., E. Kim, I. Lee, and E. M. Marcotte (2015). "Systematic comparison of variant calling pipelines using gold standard personal exome variants". In: *Sci Rep* 5, p. 17875. DOI: 10.1038/srep17875.

Ivy, R. A. et al. (2012). "Identification and characterization of psychrotolerant sporeformers associated with fluid milk production and processing". In: *Appl Environ Microbiol* 78.6, pp. 1853–64. DOI: 10.1128/AEM.06536-11.

Jackson, Brendan R. et al. (2016). "Implementation of Nationwide Real-time Whole-genome Sequencing to Enhance Listeriosis Outbreak Detection and Investigation". In: *Clinical Infectious Diseases* 63.3, pp. 380–386. DOI: 10.1093/cid/ciw242. eprint: http://oup.prod.sis.lan/cid/article-pdf/63/3/380/8039807/ciw242.pdf.

Jain, C., R. Lm Rodriguez, A. M. Phillippy, K. T. Konstantinidis, and S. Aluru (2018). "High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries". In: *Nat Commun* 9.1, p. 5114. DOI: 10.1038/s41467-018-07641-9.

Jakobsen, I. B. and S. Easteal (1996). "A program for calculating and displaying compatibility matrices as an aid in determining reticulate evolution in molecular sequences". In: *Comput Appl Biosci* 12.4, pp. 291–5.

Jimenez, Guillermo, Anicet R. Blanch, Javier Tamames, and Ramon Rossello-Mora (2013). "Complete Genome Sequence of *Bacillus toyonensis* BCT-7112T, the Active Ingredient of the Feed Additive Preparation Toyocerin". In: *Genome announcements* 1.6, e01080–13. DOI: 10.1128/genomeA.01080-13.

Joensen, K. G. et al. (2014). "Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic *Escherichia coli*". In: *J Clin Microbiol* 52.5, pp. 1501–10. DOI: 10.1128/JCM.03617-13.

Jombart, Thibaut, Michelle Kendall, Jacob Almagro-Garcia, and Caroline Colijn (2017). "treespace: Statistical Exploration of Landscapes of Phylogenetic Trees". In: *Molecular Ecology Resources* 17 (6), pp. 1385–1392.

Kabir, M. Shahjahan, Ying-Hsin Hsieh, Steven Simpson, Khalil Kerdahi, and Irshad M. Sulaiman (2017). "Evaluation of Two Standard and Two Chromogenic Selective Media for Optimal Growth and Enumeration of Isolates of 16 Unique *Bacillus* Species". In: *Journal of Food Protection* 80.6. PMID: 28467187, pp. 952–962. DOI: 10.4315/0362-028X.JFP-16-441. eprint: https://doi.org/10.4315/0362-028X.JFP-16-441.

Katz, L. S. et al. (2017). "A Comparative Analysis of the Lyve-SET Phylogenomics Pipeline for Genomic Epidemiology of Foodborne Pathogens". In: *Front Microbiol* 8, p. 375. DOI: 10.3389/fmicb.2017.00375.

Kendall, Michelle and Caroline Colijn (2015). "A tree metric using structure and length to capture distinct phylogenetic signals". In: *arXiv* 1507.05211v3. DOI: 10.1093/molbev/msw124.

— (2016). "Mapping Phylogenetic Trees to Reveal Distinct Patterns of Evolution". In: *Molecular Biology and Evolution* 33.10, pp. 2735–2743. DOI: 10.1093/molbev/msw124. eprint: http://oup.prod.sis.lan/mbe/article-pdf/33/10/2735/17472612/msw124.pdf.

Kovac, Jasna, Henk den Bakker, Laura M. Carroll, and Martin Wiedmann (2017). "Precision food safety: A systems approach to food safety facilitated by genomics tools". In: *TrAC Trends in Analytical Chemistry* 96.Supplement C, pp. 52–61.

Kovac, J. et al. (2016). "Production of hemolysin BL by *Bacillus cereus* group isolates of dairy origin is associated with whole-genome phylogenetic clade". In: *BMC Genomics* 17, p. 581. DOI: 10.1186/s12864-016-2883-z.

Kwong, J. C. et al. (2016). "Prospective Whole-Genome Sequencing Enhances National Surveillance of *Listeria monocytogenes*". In: *J Clin Microbiol* 54.2, pp. 333–42. DOI: 10.1128/JCM.02344-15.

Lewis, P. O. (2001). "A likelihood approach to estimating phylogeny from discrete morphological character data". In: *Syst Biol* 50.6, pp. 913–25.

Li, H. and R. Durbin (2010). "Fast and accurate long-read alignment with Burrows-Wheeler transform". In: *Bioinformatics* 26.5, pp. 589–95. DOI: 10.1093/bioinformatics/btp698.

Li, H., B. Handsaker, et al. (2009). "The Sequence Alignment/Map format and SAMtools". In: *Bioinformatics* 25.16, pp. 2078–9. DOI: 10.1093/bioinformatics/btp352.

Li, Heng (2013). "Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM". In: *arXiv:1303.3997v1 [q-bio.GN]*.

Liu, Y. et al. (2017). "Proposal of nine novel species of the Bacillus cereus group". In: *Int J Syst Evol Microbiol* 67.8, pp. 2499–2508. DOI: 10.1099/ijsem.0.001821.

Lotte, R. et al. (2017). "Virulence Analysis of *Bacillus cereus* Isolated after Death of Preterm Neonates, Nice, France, 2013". In: *Emerg Infect Dis* 23.5, pp. 845–848. DOI: 10.3201/eid2305.161788.

Mair-Jenkins, J. et al. (2017). "Investigation using whole genome sequencing of a prolonged restaurant outbreak of *Salmonella* Typhimurium linked to the building drainage system, England, February 2015 to March 2016". In: *Euro Surveill* 22.49. DOI: 10.2807/1560-7917.ES.2017.22.49.17-00037.

McCloskey, R. M. and A. F. Y. Poon (2017). "A model-based clustering method to detect infectious disease transmission outbreaks from sequence variation". In: *PLoS Comput Biol* 13.11, e1005868. DOI: 10.1371/journal.pcbi.1005868.

Messelhausser, U. et al. (2014). "Emetic *Bacillus cereus* are more volatile than thought: recent foodborne outbreaks and prevalence studies in Bavaria

(2007-2013)". In: *Biomed Res Int* 2014, p. 465603. DOI: 10.1155/2014/465603.

Miller, R. A., S. M. Beno, et al. (2016). "*Bacillus wiedmannii* sp. nov., a psychrotolerant and cytotoxic *Bacillus cereus* group species isolated from dairy foods and dairy environments". In: *Int J Syst Evol Microbiol* 66.11, pp. 4744–4753. DOI: 10.1099/ijsem.0.001421.

Miller, R. A., J. Jian, S. M. Beno, M. Wiedmann, and J. Kovac (2018). "Intraclade Variability in Toxin Production and Cytotoxicity of *Bacillus cereus* Group Type Strains and Dairy-Associated Isolates". In: *Appl Environ Microbiol* 84.6. DOI: 10.1128/AEM.02479-17.

Moran-Gilad, J. (2017). "Whole genome sequencing (WGS) for food-borne pathogen surveillance and control - taking the pulse". In: *Euro Surveill* 22.23. DOI: 10.2807/1560-7917.ES.2017.22.23.30547.

Morgulis, A., E. M. Gertz, A. A. Schaffer, and R. Agarwala (2006). "A fast and symmetric DUST implementation to mask low-complexity DNA sequences". In: *J Comput Biol* 13.5, pp. 1028–40. DOI: 10.1089/cmb.2006.13.1028.

Moura, A. et al. (2017). "Real-Time Whole-Genome Sequencing for Surveillance of *Listeria monocytogenes*, France". In: *Emerg Infect Dis* 23.9, pp. 1462–1470. DOI: 10.3201/eid2309.170336.

Naranjo, M. et al. (2011). "Sudden death of a young adult associated with *Bacillus cereus* food poisoning". In: *J Clin Microbiol* 49.12, pp. 4379–81. DOI: 10.1128/JCM.05129-11.

Oksanen, Jari et al. (2017). *vegan: Community Ecology Package*. R package version 2.4-2.

Olson, N. D. et al. (2015). "Best practices for evaluating single nucleotide variant calling methods for microbial genomics". In: *Front Genet* 6, p. 235. DOI: 10.3389/fgene.2015.00235.

Paradis, E., J. Claude, and K. Strimmer (2004). "APE: Analyses of Phylogenetics and Evolution in R language". In: *Bioinformatics* 20.2, pp. 289–90.

Pightling, A. W., N. Petronella, and F. Pagotto (2014). "Choice of reference sequence and assembler for alignment of *Listeria monocytogenes* short-read sequence data greatly influences rates of error in SNP analyses". In: *PLoS One* 9.8, e104579. DOI: 10.1371/journal.pone.0104579.

— (2015). "Choice of reference-guided sequence assembler and SNP caller for analysis of *Listeria monocytogenes* short-read sequence data greatly influences rates of error". In: *BMC Res Notes* 8, p. 748. DOI: 10.1186/s13104-015-1689-4.

Pruitt, K. D., T. Tatusova, and D. R. Maglott (2007). "NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins". In: *Nucleic Acids Res* 35.Database issue, pp. D61–5. DOI: 10.1093/nar/gkl842.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.

R Hackathon et al. (2019). *phylobase: Base Package for Phylogenetic Structures and Comparative Data*. R package version 0.8.6.

Revell, Liam J. (2012). "phytools: An R package for phylogenetic comparative biology (and other things)." In: *Methods in Ecology and Evolution* 3, pp. 217–223.

Rusconi, B. et al. (2016). "Whole Genome Sequencing for Genomics-Guided Investigations of *Escherichia coli* O157:H7 Outbreaks". In: *Front Microbiol* 7, p. 985. DOI: 10.3389/fmicb.2016.00985.

Sanaei-Zadeh, H. (2012). "Can *Bacillus cereus* food poisoning cause sudden death?" In: *J Clin Microbiol* 50.11, 3816, author reply 3817. DOI: 10.1128/JCM.00059-12.

Sandmann, S. et al. (2017). "Evaluating Variant Calling Tools for Non-Matched Next-Generation Sequencing Data". In: *Sci Rep* 7, p. 43169. DOI: 10.1038/srep43169.

Scallan, E. et al. (2011). "Foodborne illness acquired in the United States–major pathogens". In: *Emerg Infect Dis* 17.1, pp. 7–15. DOI: 10.3201/eid1701.P1110110.3201/eid1701.091101p1.

Schliep, Klaus, Alastair J. Potts, David A. Morrison, and Guido W. Grimm (2017). "Intertwining phylogenetic trees and networks". In: *Methods in Ecology and Evolution* 8.10, pp. 1212–1220. DOI: 10.1111/2041-210X.12760. eprint: https://besjournals.onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.12760.

Schoeni, J. L. and A. C. Wong (2005). "*Bacillus cereus* food poisoning and its toxins". In: *J Food Prot* 68.3, pp. 636–48.

Smith, J. M. (1992). "Analyzing the mosaic structure of genes". In: *J Mol Evol* 34.2, pp. 126–9.

Stamatakis, A. (2014). "RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies". In: *Bioinformatics* 30.9, pp. 1312–3. DOI: 10.1093/bioinformatics/btu033.

Stenfors Arnesen, L. P., A. Fagerlund, and P. E. Granum (2008). "From soil to gut: *Bacillus cereus* and its food poisoning toxins". In: *FEMS Microbiol Rev* 32.4, pp. 579–606. DOI: 10.1111/j.1574-6976.2008.00112.x.

Taboada, E. N., M. R. Graham, J. A. Carrico, and G. Van Domselaar (2017). "Food Safety in the Age of Next Generation Sequencing, Bioinformatics, and Open Data Access". In: *Front Microbiol* 8, p. 909. DOI: 10.3389/fmicb.2017.00909.

Tallent, S. M., K. M. Kotewicz, E. A. Strain, and R. W. Bennett (2012). "Efficient Isolation and Identification of *Bacillus cereus* Group". In: *Journal of Aoac International* 95.2, pp. 446–451. DOI: 10.5740/jaoacint.11-251.

Tallent, S. M., E. J. Rhodehamel, S. M. Harmon, and R. W. Bennett (1998). "*Bacillus cereus*". In: *Bacteriological analytical manual, 8th edition, 1998 and Foodborne pathogenic microorganisms and natural toxins handbook, 1998.* Ed. by FDA. Gaithersburg, MD: AOAC International. Chap. 14.

Taylor, Angela J. et al. (2015). "Characterization of Foodborne Outbreaks of *Salmonella enterica* Serovar Enteritidis with Whole-Genome Sequencing Single Nucleotide Polymorphism-Based Analysis for Surveillance and Outbreak Detection". In: *Journal of Clinical Microbiology* 53.10. Ed. by D. J. Diekema, pp. 3334–3340. DOI: 10.1128/JCM.01280-15. eprint: https://jcm.asm.org/content/53/10/3334.full.pdf.

Treangen, T. J., B. D. Ondov, S. Koren, and A. M. Phillippy (2014). "The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes". In: *Genome Biol* 15.11, p. 524. DOI: 10.1186/PREACCEPT-2573980311437212.

Turnbull, P. C. and J. M. Kramer (1985). "Intestinal carriage of *Bacillus cereus*: faecal isolation studies in three population groups". In: *J Hyg (Lond)* 95.3, pp. 629–38.

Usongo, V. et al. (2018). "Impact of the choice of reference genome on the ability of the core genome SNV methodology to distinguish strains of *Salmonella enterica* serovar Heidelberg". In: *PLoS One* 13.2, e0192233. DOI: 10.1371/journal.pone.0192233.

Vangay, P., E. B. Fugett, Q. Sun, and M. Wiedmann (2013). "Food microbe tracker: a web-based tool for storage and comparison of food-associated microbes". In: *J Food Prot* 76.2, pp. 283–94. DOI: 10.4315/0362-028X.JFP-12-276.

Walker, T. M. et al. (2018). "A cluster of multidrug-resistant *Mycobacterium tuberculosis* among patients arriving in Europe from the Horn of Africa: a molecular epidemiological study". In: *Lancet Infect Dis* 18.4, pp. 431–440. DOI: 10.1016/S1473-3099(18)30004-5.

Wickham, Hadley (2017). *stringr: Simple, Consistent Wrappers for Common String Operations*. R package version 1.2.0.

Ye, J., S. McGinnis, and T. L. Madden (2006). "BLAST: improvements for better sequence analysis". In: *Nucleic Acids Res* 34.Web Server issue, W6–9. DOI: 10.1093/nar/gkl164.

Yu, Guangchuang, David K. Smith, Huachen Zhu, Yi Guan, and Tommy Tsan-Yuk Lam (2017). "ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data". In: *Methods in Ecology and Evolution* 8.1, pp. 28–36. DOI: doi:10.1111/2041-210X.12628.

CHAPTER 6

CONCLUSION

Foodborne disease-causing agents have been estimated to cause more than 600 million illnesses and more than 400,000 deaths worldwide annually (WHO 2015). Due to their profound human and economic impact, there is incentive to query bacterial disease agents responsible for a significant proportion of illnesses, deaths, and disease burden using whole-genome sequencing (WGS); congruent with this, the amount of publicly available sequencing data derived from microbes has doubled in size every two years, and will likely continue to grow increasingly (Bradley, Bakker, et al. 2019). The previous chapters detail how Illumina sequencing data from thousands of bacterial isolates can be leveraged to draw meaningful biological conclusions relevant to food safety and quality.

6.1 NGS can be used to replicate many microbiological assays *in silico* with high accuracy, speed, and throughput

As demonstrated in Chapters 2 (L. M. Carroll, M. Wiedmann, et al. 2017) and 4 (L. M. Carroll, Kovac, et al. 2017), numerous assays used to characterize food-associated microorganisms can be replicated *in silico* using NGS, often with the advantage of increased speed and throughput. In Chapter 2, whole-genome sequencing (WGS) was used to query *Salmonella enterica* serotypes capable of infecting both bovine and human hosts (i.e., serotypes Dublin, Newport, and Typhimurium) from bovine and human sources in different geographic regions of the United States (New York State on the east coast, and Washington State

on the west coast). *In silico* detection of antimicrobial resistance (AMR) determinants was able to predict phenotypic resistance to antimicrobials used in human and veterinary medicine with high accuracy (L. M. Carroll, M. Wiedmann, et al. 2017). Additionally, *in silico Salmonella* serotype designations were consistent with (and, sometimes even more accurate than) those assigned using traditional serotyping (L. M. Carroll, M. Wiedmann, et al. 2017). These results further support that WGS can be used to reliably predict AMR phenotypes and *Salmonella* serotype (Bradley, Gordon, et al. 2015; McDermott et al. 2016; S. Zhang et al. 2015; Yoshida et al. 2016) and attest to the robustness of these *in silico* assays in not only human clinical isolates, but those of animal (i.e., bovine) origin as well (L. M. Carroll, M. Wiedmann, et al. 2017).

In Chapter 4 (L. M. Carroll, Kovac, et al. 2017), PCR-based detection of virulence factors, as well as single- (i.e., *panC* and *rpoB*) and multi-locus sequence typing for multiple species in the *Bacillus cereus* group, were shown to be readily replicated *in silico* with high accuracy. Additionally, when implemented in a freely available and open-source pipeline, these *in silico* assays could be scaled to hundreds of genomes to gain insight into the population structure and virulence capacity of all known members of the *B. cereus* group. While efforts to sequence *B. cereus* strains are not as well-established as those for other foodborne pathogens (e.g., *Salmonella enterica*, *Listeria monocytogenes*), the number of publicly available *B. cereus* group genomes is increasing (Laura M. Carroll, Martin Wiedmann, et al. 2019). As such, scalable, rapid *in silico* typing methods will become increasingly valuable and will offer further insight into the genomics of the group, with the potential to explore novel lineages important to food safety, quality, and human health (e.g., as was done for proposed novel *B. cereus* group species "*Bacillus clarus*") (Acevedo et al. 2019).

6.2 NGS can be used to identify novel genomic elements associated with clinically relevant phenotypes

In addition to replicating existing microbiological assays, NGS can be used to identify novel associations between genomic elements and phenotypes of interest, as was demonstrated in Chapter 3 (Laura M. Carroll, Gaballa, et al. 2019): during routine *in silico* screening of sequenced *Salmonella enterica* genomes, a novel mobilized colistin resistance gene, *mcr-9*, was identified based on its similarity to existing *mcr* homologues (Laura M. Carroll, Gaballa, et al. 2019). While *mcr-9* was confirmed to confer resistance to colistin up to and beyond the clinical breakpoint when cloned into *Escherichia coli*, the *Salmonella* Typhimurium isolate in which it was initially detected was not itself clinically resistant (Laura M. Carroll, Gaballa, et al. 2019). This approach can be contrasted with the "traditional" approach to *mcr* identification, in which a colistin-resistant bacterial isolate is used to identify *mcr* homologues, as was done to identify *mcr-1, -2, -3, -4, -5, -7*, and *-8* (Liu et al. 2016; Xavier et al. 2016; Yin et al. 2017; Carattoli et al. 2017; Borowiak et al. 2017; Yang et al. 2018; Wang et al. 2018) (in the case of *mcr-6*, a colistin-sensitive *Moraxella* strain was screened for *mcr-1* and *mcr-2* and was found to harbor a *mcr-2*-like gene, which was later renamed *mcr-6*) (AbuOun et al. 2017; Partridge et al. 2018). In the case of *mcr-9*, the traditional route of *mcr* identification (i.e., testing for bacterial resistance to colistin, and then identifying *mcr*-like genes if the isolate is colistin-resistant at the clinical breakpoint under standard testing conditions) would have left it undetected. It is likely that routine *in silico* screening of *Enterobactericiae* genomes will yield other *mcr* genes capable of conferring resistance to colistin. However, as was the case with *mcr-9*, future studies to determine the conditions under which differ-

ent *mcr* homologues are transcribed and expressed are warranted. Furthermore, the current view of colistin resistance (and antimicrobial resistance as a whole), strictly though the lens of a susceptible-resistant dichotomy, warrants critique, as testing conditions have been shown to influence *mcr* expression and colistin minimum inhibitory concentration (MIC) (H. Zhang et al. 2017; Gwozdzinski et al. 2018).

6.3 NGS can be used to query pathogens associated with foodborne outbreaks at higher resolution than its predecessors

NGS technologies have been implemented in public health settings to routinely sequence numerous foodborne pathogens, including *Salmonella enterica, Listeria monocytogenes,* and *Escherichia coli* (Taylor et al. 2015; Hoffmann et al. 2016; Gymoese et al. 2017; Grad et al. 2012; Holmes et al. 2015; Rusconi et al. 2016; Jackson et al. 2016; Kwong et al. 2016; Chen, Luo, Pettengill, et al. 2017; Chen, Luo, Curry, et al. 2017; Moura et al. 2017). Chapter 5 offered the first description of a foodborne outbreak caused by members of the *Bacillus cereus* group in which WGS was used to characterize isolates (Laura M. Carroll, Martin Wiedmann, et al. 2019). In addition to providing the level of expected diversity among emetic *Bacillus cereus* outbreak isolates obtained via different variant calling methodologies, the study presented in Chapter 5 showcases that WGS can reliably differentiate emetic *B. cereus* strains from a single-source outbreak from publicly available genomes of the same sequence type and virulotype, even in the absence of large amounts of genomic data from *B. cereus* group genomes (Laura M. Carroll, Martin Wiedmann, et al. 2019). Additionally, the

value (or lack thereof) of various metrics which might serve as supplemental metadata (e.g., patient symptoms, bacterial counts) were discussed; in the outbreak presented here, cytotoxicity data proved to be particularly useful for excluding non-emetic *Bacillus cereus* group isolates from the outbreak, and, thus, the possibility of a multi-source outbreak caused by multiple species (Laura M. Carroll, Martin Wiedmann, et al. 2019). The computational, microbiological, and epidemiological methods presented in this study will benefit not only *Bacillus cereus* researchers, but also those in public health who are working with under-studied and under-reported pathogens, particularly those which may be ubiquitous in the environment or varying in their virulence capacity.

Overall, NGS technologies are being used increasingly in food safety and public health settings, with the advantage of not only replicating microbiological assays *in silico*, but providing opportunities to develop novel bacterial characterization schemes which query the genomes of bacterial pathogens in their entirety. Decreasing sequencing costs and increasingly available genomic data from food-associated microbes and communities will allow for improved biological inference from farm to fork.

6.4 References

AbuOun, M. et al. (2017). "*mcr-1* and *mcr-2* variant genes identified in *Moraxella* species isolated from pigs in Great Britain from 2014 to 2015". In: *J Antimicrob Chemother* 72.10, pp. 2745–2749. DOI: 10.1093/jac/dkx286.

Acevedo, Marysabel Mendez et al. (2019). "*Bacillus clarus* sp. nov. is a new *Bacillus cereus* group species isolated from soil". In: *bioRxiv*. DOI: 10.1101/508077. eprint: https://www.biorxiv.org/content/early/2019/01/02/508077.full.pdf.

Borowiak, M. et al. (2017). "Identification of a novel transposon-associated phosphoethanolamine transferase gene, *mcr-5*, conferring colistin resistance in d-tartrate fermenting *Salmonella enterica* subsp. enterica serovar Paratyphi B". In: *J Antimicrob Chemother* 72.12, pp. 3317–3324. DOI: 10.1093/jac/dkx327.

Bradley, Phelim, Henk C. den Bakker, Eduardo P. C. Rocha, Gil McVean, and Zamin Iqbal (2019). "Ultrafast search of all deposited bacterial and viral genomic data". In: *Nature Biotechnology* 37.2, pp. 152–159. DOI: 10.1038/s41587-018-0010-1.

Bradley, Phelim, N. Claire Gordon, et al. (2015). "Rapid antibiotic-resistance predictions from genome sequence data for *Staphylococcus aureus* and *Mycobacterium tuberculosis*". In: *Nat Commun* 6, pp. 10063–10063. DOI: 10.1038/ncomms10063.

Carattoli, A. et al. (2017). "Novel plasmid-mediated colistin resistance *mcr-4* gene in *Salmonella* and *Escherichia coli*, Italy 2013, Spain and Belgium, 2015 to 2016". In: *Euro Surveill* 22.31. DOI: 10.2807/1560-7917.ES.2017.22.31.30589.

Carroll, L. M., J. Kovac, R. A. Miller, and M. Wiedmann (2017). "Rapid, high-throughput identification of anthrax-causing and emetic *Bacillus cereus* group genome assemblies using BTyper, a computational tool for virulence-based classification of *Bacillus cereus* group isolates using nucleotide sequencing data". In: *Appl Environ Microbiol*. DOI: 10.1128/AEM.01096-17.

Carroll, L. M., M. Wiedmann, et al. (2017). "Whole-Genome Sequencing of Drug-Resistant *Salmonella enterica* Isolates from Dairy Cattle and Humans in New York and Washington States Reveals Source and Geographic Associations". In: *Appl Environ Microbiol* 83.12. DOI: 10.1128/AEM.00140-17.

Carroll, Laura M., Ahmed Gaballa, et al. (2019). "Identification of Novel Mobilized Colistin Resistance Gene *mcr-9* in a Multidrug-Resistant, Colistin-Susceptible *Salmonella enterica* Serotype Typhimurium Isolate". In: *mBio* 10.3. Ed. by Mark S. Turner, Gregory Siragusa, and David White. DOI: 10.1128/mBio.00853-19. eprint: https://mbio.asm.org/content/10/3/e00853-19.full.pdf.

Carroll, Laura M., Martin Wiedmann, et al. (2019). "Characterization of Emetic and Diarrheal *Bacillus cereus* Strains From a 2016 Foodborne Outbreak Using Whole-Genome Sequencing: Addressing the Microbiological, Epidemiological, and Bioinformatic Challenges". In: *Frontiers in Microbiology* 10.144. DOI: 10.3389/fmicb.2019.00144.

Chen, Y., Y. Luo, P. Curry, et al. (2017). "Assessing the genome level diversity of *Listeria monocytogenes* from contaminated ice cream and environmental samples linked to a listeriosis outbreak in the United States". In: *PLoS One* 12.2, e0171389. DOI: 10.1371/journal.pone.0171389.

Chen, Y., Y. Luo, J. Pettengill, et al. (2017). "Singleton Sequence Type 382, an Emerging Clonal Group of *Listeria monocytogenes* Associated with Three Multistate Outbreaks Linked to Contaminated Stone Fruit, Caramel Apples, and Leafy Green Salad". In: *J Clin Microbiol* 55.3, pp. 931–941. DOI: 10.1128/JCM.02140-16.

Grad, Y. H. et al. (2012). "Genomic epidemiology of the *Escherichia coli* O104:H4 outbreaks in Europe, 2011". In: *Proc Natl Acad Sci U S A* 109.8, pp. 3065–70. DOI: 10.1073/pnas.1121491109.

Gwozdzinski, K., S. Azarderakhsh, C. Imirzalioglu, L. Falgenhauer, and T. Chakraborty (2018). "An Improved Medium for Colistin Susceptibility Testing". In: *J Clin Microbiol* 56.5. DOI: 10.1128/JCM.01950-17.

Gymoese, P. et al. (2017). "Investigation of Outbreaks of *Salmonella enterica* Serovar Typhimurium and Its Monophasic Variants Using Whole-Genome Sequencing, Denmark". In: *Emerg Infect Dis* 23.10, pp. 1631–1639. DOI: 10.3201/eid2310.161248.

Hoffmann, M. et al. (2016). "Tracing Origins of the *Salmonella* Bareilly Strain Causing a Food-borne Outbreak in the United States". In: *J Infect Dis* 213.4, pp. 502–8. DOI: 10.1093/infdis/jiv297.

Holmes, A. et al. (2015). "Utility of Whole-Genome Sequencing of *Escherichia coli* O157 for Outbreak Detection and Epidemiological Surveillance". In: *J Clin Microbiol* 53.11, pp. 3565–73. DOI: 10.1128/JCM.01066-15.

Jackson, Brendan R. et al. (2016). "Implementation of Nationwide Real-time Whole-genome Sequencing to Enhance Listeriosis Outbreak Detection and Investigation". In: *Clinical Infectious Diseases* 63.3, pp. 380–386. DOI: 10.1093/cid/ciw242. eprint: http://oup.prod.sis.lan/cid/article-pdf/63/3/380/8039807/ciw242.pdf.

Kwong, J. C. et al. (2016). "Prospective Whole-Genome Sequencing Enhances National Surveillance of *Listeria monocytogenes*". In: *J Clin Microbiol* 54.2, pp. 333–42. DOI: 10.1128/JCM.02344-15.

Liu, Y. Y. et al. (2016). "Emergence of plasmid-mediated colistin resistance mechanism *MCR-1* in animals and human beings in China: a microbiological and molecular biological study". In: *Lancet Infect Dis* 16.2, pp. 161–8. DOI: 10.1016/S1473-3099(15)00424-7.

McDermott, Patrick F. et al. (2016). "Whole-Genome Sequencing for Detecting Antimicrobial Resistance in Nontyphoidal *Salmonella*". In: *Antimicrobial Agents and Chemotherapy* 60.9, pp. 5515–5520. DOI: 10.1128/AAC.01030-16. eprint: https://aac.asm.org/content/60/9/5515.full.pdf.

Moura, A. et al. (2017). "Real-Time Whole-Genome Sequencing for Surveillance of *Listeria monocytogenes*, France". In: *Emerg Infect Dis* 23.9, pp. 1462–1470. DOI: 10.3201/eid2309.170336.

Partridge, S. R. et al. (2018). "Proposal for assignment of allele numbers for mobile colistin resistance (*mcr*) genes". In: *J Antimicrob Chemother* 73.10, pp. 2625–2630. DOI: 10.1093/jac/dky262.

Rusconi, B. et al. (2016). "Whole Genome Sequencing for Genomics-Guided Investigations of *Escherichia coli* O157:H7 Outbreaks". In: *Front Microbiol* 7, p. 985. DOI: 10.3389/fmicb.2016.00985.

Taylor, Angela J. et al. (2015). "Characterization of Foodborne Outbreaks of *Salmonella enterica* Serovar Enteritidis with Whole-Genome Sequencing Single Nucleotide Polymorphism-Based Analysis for Surveillance and Outbreak Detection". In: *Journal of Clinical Microbiology* 53.10. Ed. by D. J. Diekema, pp. 3334–3340. DOI: 10.1128/JCM.01280-15. eprint: https://jcm.asm.org/content/53/10/3334.full.pdf.

Wang, X. et al. (2018). "Emergence of a novel mobile colistin resistance gene, *mcr-8*, in NDM-producing *Klebsiella pneumoniae*". In: *Emerg Microbes Infect* 7.1, p. 122. DOI: 10.1038/s41426-018-0124-z.

WHO (2015). *WHO estimates of the global burden of foodborne diseases, 2007-2015*. WHO, Geneva, Switzerland.

Xavier, B. B. et al. (2016). "Identification of a novel plasmid-mediated colistin-resistance gene, *mcr-2*, in *Escherichia coli*, Belgium, June 2016". In: *Euro Surveill* 21.27. DOI: 10.2807/1560-7917.ES.2016.21.27.30280.

Yang, Y. Q., Y. X. Li, C. W. Lei, A. Y. Zhang, and H. N. Wang (2018). "Novel plasmid-mediated colistin resistance gene *mcr-7.1* in *Klebsiella pneumoniae*". In: *J Antimicrob Chemother*. DOI: 10.1093/jac/dky111.

Yin, W. et al. (2017). "Novel Plasmid-Mediated Colistin Resistance Gene *mcr-3* in *Escherichia coli*". In: *MBio* 8.3. DOI: 10.1128/mBio.00543-17.

Yoshida, C. E. et al. (2016). "The *Salmonella In Silico* Typing Resource (SISTR): An Open Web-Accessible Tool for Rapidly Typing and Subtyping Draft *Salmonella* Genome Assemblies". In: *PLoS One* 11.1, e0147101. DOI: 10.1371/journal.pone.0147101.

Zhang, H. et al. (2017). "Expression characteristics of the plasmid-borne *mcr-1* colistin resistance gene". In: *Oncotarget* 8.64, pp. 107596–107602. DOI: 10.18632/oncotarget.22538.

Zhang, S. et al. (2015). "*Salmonella* serotype determination utilizing high-throughput genome sequencing data". In: *J Clin Microbiol* 53.5, pp. 1685–92. DOI: 10.1128/JCM.00323-15.